Combining Self Learning and Active Learning for Chinese Named Entity Recognition*

Lin Yao[1], Chengjie Sun[2], Xiaolong Wang[1], Xuan Wang[1]
[1]Computer Science Department, Harbin Institute of Technology Shenzhen Graduate School
ShenZhen, China
[2] School of Computer Science and Technology, Harbin Institute of Technology
Harbin, China
{yaolin, cjsun,Wangxl, wangxuan}@insun.hit.edu.cn

Abstract—This paper proposes a combination of active learning and self-training method to reduce the labeling effort for Chinese Named Entity Recognition (NER). Active learning and self-training are two different ways to use unlabeled data. They are complement when choosing unlabeled data for further training. A new strategy based on Information Density (ID) for sample selecting in sequential labeling problem is also proposed, which is suitable for both active learning and self-training. Conditional Random Fields (CRFs) is chosen as the underlying model for active learning and self-training in the proposed approach due to its promising performance in many sequence labeling tasks. Experiment results show the effect of the proposed method. On Sighan bakeoff 2006 MSRA NER corpus, an F1 score of 77.4% is achieved by using only 15,000 training sentences chosen by the proposed hybrid method.

***Index Terms*—self-training, active learning, named entity recognition, information density, condition random fields**

I. INTRODUCTION

Named Entity Recognition (NER), subtask of information extraction, locates atomic elements from text and labels them with pre-defined categories such as person name, locations, organizations etc. Due to its important roles in many NLP applications, NER have become the share task of MUC-6, MUC-7, Conll2002 and Conll2003. Several years endeavor of many researchers flourish the development of this area. Numerous different machine learning methods such as Maximum Entropy[1-4], Hidden Markov Models[5-8], Support Vector Machines [7, 9] and Conditional Random Fields [10, 11] have been adopted for NER and achieved high accuracy. Most of these methods are based on supervised machine learning. The better predictive ability is achieved based on large and complex data which try to represent as much as possibility. However, generally even thousands of labeled examples can only sparsely cover the parameter space.[12]. Moreover, in sequence modeling like NER task, it is more difficult to obtain labeled training data since hand-labeling individual words and word boundaries is really complex and need professional annotators. Therefore, the shortage of annotated training sets is the obstacle of supervised learning and limits the further development, especially for some languages which are not well know. To reduce the number of labeled training samples to shorten the training time and to achieve the requested performance is not a trivial task.

Many new methods are studied to solve the shortage of labeled data. Active learning (AL) is the method which, instead of relying on random sampling from the large training data, actively participates in the optimal choosing training examples. Using different strategies, active learning may determine much smaller and most informative subset from the unlabeled data pool. In addition, active learning with the help of other techniques can construct its own training data and achieve satisfied results.

Manually annotations are somehow expensive. This situation gives researchers motivation to exploit the remaining data set which is not labeled by a human. The semi-supervised learning way is explored to solve the task. The attractiveness of semi supervised learning for language tasks is unlabeled data is so plentiful and considerably cheaper to obtain compared with hand-labeling data.

In this paper, we present an alternative active learning strategy, and combine this method with semi-supervised learning for the NER task. Without large-scale labeled data, the proposed method greatly reduces the training efforts and gets similar even better results as compared to the conventional supervised learning mode.

The organization of the paper is as follows. Following the introduction in Section I, Section II presents the related works briefly. In Section III an introduction of associated knowledge is given. It includes the general CRFs framework and the description of active learning model, followed by our new query strategy based on information density. The semi-supervised learning is also presented in this Section. Then in the next Section the hybrid model of active learning and semi-supervised is proposed. We describe system procedure of our system in Section V. We present and analyze our results in the following Section. Finally, some concluding remarks are presented in Section VII.

II. RELATED WORK

The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with

* This investigation was supported by the project of the High Technology Research and Development Program of China (grant No. 2006AA01Z197, 2007AA01Z194), the project of the National Natural Science Foundation of China (grant No. 60673037)

doi:10.4304/jsw.5.5.530-537

fewer labeled training instances if it is allowed to choose the data from which is learns. One of the first active learning scenarios to be investigated is learning with membership queries [13]. Since that, active learning was applied to different supervised learning models, such as support-vector machines [14, 15], conditional random field[16], as well as Boosting and Bagging [17]. The active learning method has been studied for many real-world problem domains in machine learning, such as text classification[15, 18-20] , information extraction[21-23], word segmentation [24], image classification and retrieval[25, 26] , video classification and retrieval[27, 28], spoken language understanding[29], automatic speech Recognition[30], and cancer diagnosis [31].

Semi-supervised learning algorithms is another method which based on a small amount labeled data, applies a large number of unlabeled data to reduce the dependency on labeled training data. Self-training and Co-training are two commonly used techniques for semi-supervised learning. Yarowsky [32] applies self-training to word sense disambiguation. Nigam and Ghani [33] compare co-training with generative mixture models and EM. Self-training is used by Riloff et al.[34] to identify subjective nouns. Rosenberg et al. use self-training for object detection systems from images [35]. Collins and Singer [36] , Jones [37] used co-training, co-EM and other related methods for information extraction from text. We select self-training as the semi-supervised learning methods used in our system.

The first mix model of active and semi-supervised learning was applied to assign class labels to unlabeled examples and introduced by McCallum and Nigam [38]. They combined the active learning algorithm with EM-style semi-supervised learning to assign class labels to those examples that remain unlabeled. This idea was developed. Muslea et al. [39] introduce a new multi-view algorithm, Co-EMT, which combines semi-supervised and active learning for text classification. Exploiting multiple views for both active and semi-supervised learning has been shown to be very effective. The combination of active and semi-supervised learning was used for automatic speech recognition and have shown improvements for statistical language modeling [40].

III. ASSOCIATED KNOWLEDGE

A. Conditional random field

Similar to the Hidden Markov Models (HMMs)[41], Conditional Random Fields (CRFs)[42] are a probabilistic framework for labeling and segmenting sequential data. A conditional Random fields is an undirected graphical model and calculates the conditional probability of output values based on given input values. To reduce complexity, strong independence assumption is made between observation variables when HMMs is used, which impairs the accuracy of the model. When using CRF, it does not need to make assumptions on the dependencies among observation variables, which is different from HMMs. The CRFs have been applied in many domains to deal with the structured data. Because of its linear structure, Linear-chain CRFs is frequently chosen to deal with the linear labeling questions. Fig. 1 shows the graphical representation of liner-chain CRFs.

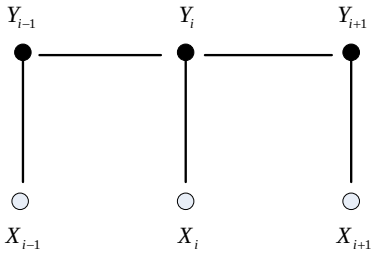

Figure 1. Graphical representation of liner-chain CRFs

A linear-chain CRFs model is described as following:

$$P\left(\vec{y}/\vec{x}\right)=\frac{1}{Z_{\vec{x}}}\exp\left(\sum_{t=1}^{n}\sum_{k}\lambda_k f_k\left(y_{t-1},y_t,\vec{x},t\right)\right) \quad (1).$$

With the observation sequence $x=\left\langle x_1,x_2,...,x_n\right\rangle$ and the label sequence $y=\left\langle y_1,y_2,...,y_n\right\rangle$.

In equation (1), y_t is the label for position t, state feature function is concerned y_t with the entire observation sequence, the transition feature between labels of position t-1 and t on the observation sequence is also considered [43]. Each feature f_k can be either state feature function or transition feature function. λs are the parameters to weight corresponding features and can be estimated from training data. Z is a normalization factor.

B. Active learning

Using the supervised learning methods such as CRFs, HMMs and Maximum Entropy Markov Model (MEMM), large number of labeled data are required for training these models. Labeling-required training data is a time-consuming job, Zhu's report [44] shows that one minute of speech takes ten minutes to label at the word level, and to annotate phonemes can take nearly seven hours, which is 400 times longer than original speech. Labeling-required training data is also an expensive task, requires many trained annotators, in some areas, such as biomedical information extraction even require PhD-level biologists to label the data. Reduction of the dependence on large amount labeled training data relies on great growth of learning ability. According to the algorithm 1, active learning is a solution for the problem with scarce labeled data and rich unlabeled data in supervised learning.

In active learning, the training procedure begins with a small number labeled training set. Thereafter, according to a particular query strategy, the most informative instance X is selected from unlabeled data pool U, and labeled by human annotators. Labeled X is added into previous training set and the training procedure remains continue. The iteration becomes halt when the stopping criterion is met. The active data selection is expected to improve the system accuracy compared with the random data selection.

1) Active Learning Scenarios

Based on different active learning settings, there are 3 main types active learning [45]: (i) membership query synthesis, (ii) stream-based selective sampling, and (iii) pool-based active learning.

First, generally, the membership query synthesis is used in regression learning tasks. It requests every unlabeled instance in the input space to be labeled without concerning the potential natural distribution and may cause awkward amount of work for human annotator. Latter, two scenarios are provided to address these limitations.

Algorithm 1 Pool-based active learning

Input: training set L, pooled Unlabeled data U

Output: model θ

While not meet the stopping criteria

//training a model using training set L

$\theta = train(L)$

//determining the most informativeness subset based on the strategy formula Φ(x) from the unlabeled data set. S is the fetch size

$X_{(s)} = \arg\max_{x \in U} \varphi(x)$

// Add selected subset into training set L

$L = L \cup label(X_s)$

$U = U - X_s$

Second, for the stream-based selective sampling, unlabeled data flows out from the data source sequentially quite like steam. It can be queried or discarded by the learner based on individual instance.

Instead of relying on single data instance, the third pool-based active learning, which is used by more learners, ranks the entire data and choose the most valuable subset from the data pool. Pool-based active learning is extremely suitable for the learning task with strict memory, processing power requirement and abundant unlabeled raw data.

The supervised machine learning methods for NER usually rely on a large amount of labeled sentences. There are two disadvantages of these methods. First, no matter which method is chosen, the training procedure is time consuming. Second, we cannot add new training samples one by one into the previous labeled data set due to the problem of large consumption of time. Pool-based active learning allows the learner to query instances in groups, which is better suited to parallel labeling environments or models with slow training procedure. Thus the pool-based active learning is selected and used in our NER system.

2) Active Learning Query Strategies

Active Learning methods rely on different strategies for sampling unlabeled instances. There have been many proposed methods of formulating such query strategies in the literature. In the following subsection, we discuss several existing representative learning methods and propose the strategy, which is used in our system.

Generally, the algorithm for query strategy used frequently is uncertainty sampling, which queries the least certain instances from the unlabeled data pool. Query-by-committee is another algorithm, using different committee members and tactical voting to pick up the subset which is most disagreed by the committee. The third approach is expected model change, which prefers the instances influence the model most effectively.

3) Information Density Strategy

Prior algorithms have their limitations. As depicted in Fig. 2, the most uncertain instances node is A, so uncertainty sampling methods will label A, and add it into the labeled instance set L. However, B is closer to other unlabeled instances and should have more representative information than A. Therefore, B should be chosen instead of A. Information density can take advantage of the average similarity between the target instance and other unlabeled instances to abstain the prior wrong sampling. Moreover, the similarity between the target instance and labeled instances should be considered. When compared with C, instance B is close to the labeled node D, which means B is similar to D and node C has more informativeness and should be queried instead of B. To address these issues, we propose a modified information density query strategy, which is formulated as (2).

$$\phi^{ID}(x) = \phi^{SE}(x) \times \left(\frac{1}{U} \sum_{u=1}^{U} Sim(x, x^u) \right)^{\beta} \times e^{-sim(x.x^L)} \quad (2)$$

$$\phi^{SE}(x) = -\sum_{\hat{y}} P(\hat{y} \mid x; \theta) \log P(\hat{y} \mid x; \theta) \quad (3)$$

Sequence entropy ϕ^{SE} is used to evaluate the basic informativeness of current instance; $\hat{y}$ ranges over all possible label sequences with the input sequence x. The larger value of ϕ^{SE} means the node has the larger uncertainty and more useful information for the system. The average distance between current sequence and unlabeled pooled data is presented by $\left(\frac{1}{U} \sum_{u=1}^{U} Sim(x, x^u) \right)^{\beta}$ which is investigated based on Cosine similarity function. The similarity with labeled data is shown as $e^{-sim(x.x^L)}$.

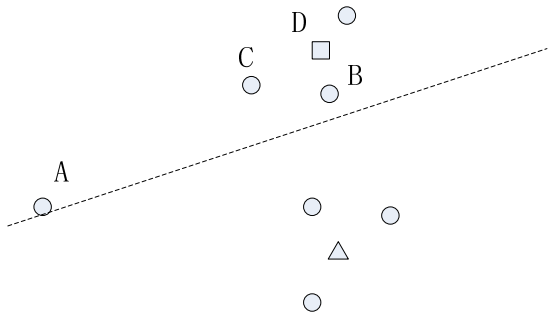

Figure 2. Analysis for uncertainty sampling

The sequence of feature vectors X, is represented by the matrix as shown in Figure 3. Each token in a sequence is described by a feature vector $(f_1 \ldots f_J)$; J is the number of features. To calculate the similarity, the sequence of feature vectors X is transform into a single vector $\vec{x}$ as shown in (4).

$$\vec{x} = \left[\sum_{t=1}^{T} f_1(x_t), \ldots, \sum_{t=1}^{T} f_J(x_t) \right] \quad (4)$$

sequence	f_1	f_2	. . .	f_J
$token_1$	x_{11}	x_{12}	. . .	x_{1J}
$token_2$	x_{21}	x_{22}	. . .	x_{2J}
⋮	⋮			⋮
$token_T$	x_{T1}	x_{T2}	. . .	x_{TJ}

Figure 3. Matrix of sequence feature vectors

In (4), the summation $\sum_{t=1}^{T} f_j(x_t)$ stands for the sum of f_j column across all tokens. However, the value x_{ij} could be nominal value and cannot be simply added together. The nominal values should be casted into numeric ones. The new vectors are assigned to the nominal values. For example, the Part of Speech (PoS) feature is mapped to the binary value vector. The dimension of vector maps the number of PoS tags. If the PoS tag is NN, then the dimension of NN will be assign to 1 and others will be 0.

Fig.4 presents the procedure of active learning procedure. The kernel part of this procedure is the sample selection using information density strategy. Similar to Sequence Entropy algorithm, we are looking for the instances with most uncertainty. Intuitively, these samples cannot be labeled correctly but is informativeness for improving the system. Theoretically mapping to the graph, these points are distributed around the decision boundary. Adding such labeled samples improves the generosity of the trained models more effectively than random pickup.

C. Semi-supervise Learning

1) Semi-supervise learning…

Supervised learning is a machine learning technique for learning a model from a huge number of training data. Generally the supervised learning can provide better performance since it is rely on the labeled data. However, the labeled instances are difficult and expensive to collect. Unsupervised learning is distinguished from supervised learning in that the learner is given only unlabeled examples, which is relatively easy to obtain. Nevertheless, it loses the accuracy. Semi-supervised learning is halfway between supervised and unsupervised learning and makes use of both labeled and unlabeled data for training. Because of the less human effort but higher precision, semi-supervised is interested in many areas.

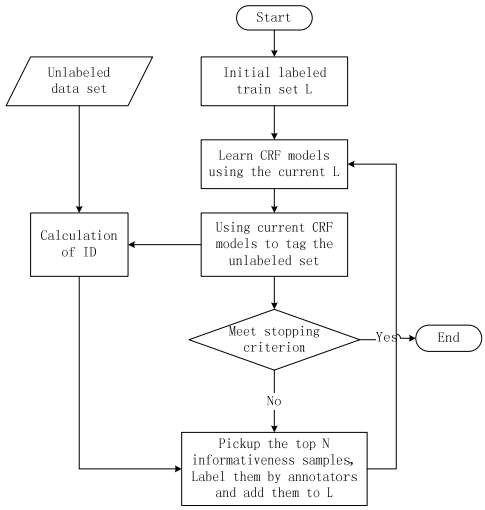

Figure 4. Active learning procedure

2) Self-training

Self-training is one of semi-supervised learning methods and has been widely used in NLP research. In self-training, first, a small amount of labeled data is used, like the supervised method, and an underlying classifier is obtained. Then the classifier is used to label the unlabeled data. According to the confident score, the unlabeled instances which have higher credibility than the threshold are added into the training set. The earlier classifier is retrained and this procedure is repeated until the stop criterion is met. There is no perfect selection. The choosing threshold is about tradeoffs. The higher threshold means less useful informative data, while lower value of threshold means noisy. The self-training procedure is described as Fig. 5.

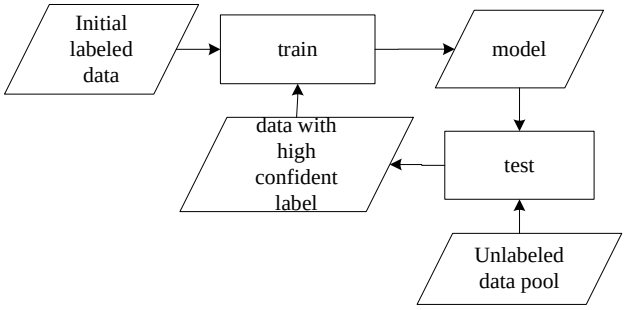

Figure 5. The procedure of self learning

IV. HYBRID OF ACTIVE LEARNING AND SEMISUPERVISED LEARNING

Both Semi-supervised learning and active learning

reduce the annotation effort by learning from small annotated data. In active learning, the group of unlabeled data with lowest confident scores is believed to have more information for the underlying model to learn and these data will be chosen and join into the later retraining. How to exploit the remaining set of data with higher confident score. Semi-supervised learning is the answer. In semi-supervised learning, the higher confident scores means the credibility and these data are used for later procedure. Semi-supervised learning only chooses the data which are classified with confidence scores higher than some threshold. This action is the opposite to the active learning. The complement selection is achieved jointly by active learning and semi-supervised learning.

For the existing semi-supervised approaches, since the initial human labeled data is limited, the poor training could cause the early mistake and these mistakes reinforce themselves. Combining active learning with semi-supervised learning may be a solution to this problem.

The algorithm 2 describes the combination of active learning presented in Section 4 and self-training presented in Section 5. In the hybrid learning method, the training set is initialized by labeled data set L. The classifier is trained using the training set T. Instead of fix numbered, the training set is dynamic increasing by adding the human-labeled data and machine-labeled data. The generated model joins into the next round classification until the stopping criterion is met. The human-labeled data is selected based on active learning method and the self-training generates the machine-labeled data.

After each round classification, the labeled data is selected to be relabeled by human annotator or added into the training set directly based on the threshold. Note that in this algorithm, the mixture of human-labeled data and machine-labeled data are added into the training set and join the retraining from the first round. A simplified flowchart is depicted in Fig. 6.

Algorithm 2 the combination of active learning and semi-supervised learning

Input: human-labeled data set L, Unlabeled data set U

Initialize the training set T = L

While *not meet the stopping criteria*

//training a model using training set T

$\theta = train(T)$

//determining the most informativeness subset based on the strategy formula φ(x) from the unlabeled data set. S is the fetch size

$X_{(S)} = \arg\max_{x \in U} \varphi(x)$

//determining the most confidently labeled subset based on the strategy formula φ(y) from the unlabeled data set. N is the fetch size

$Y_{(N)} = \arg\min_{y \in U} \varphi(y)$

// Add selected subset into training set T

$T = T \cup human-labeled(X_S) \cup machine-labeled(Y_N)$

$U = U - X_S - Y_N$

Output: model θ

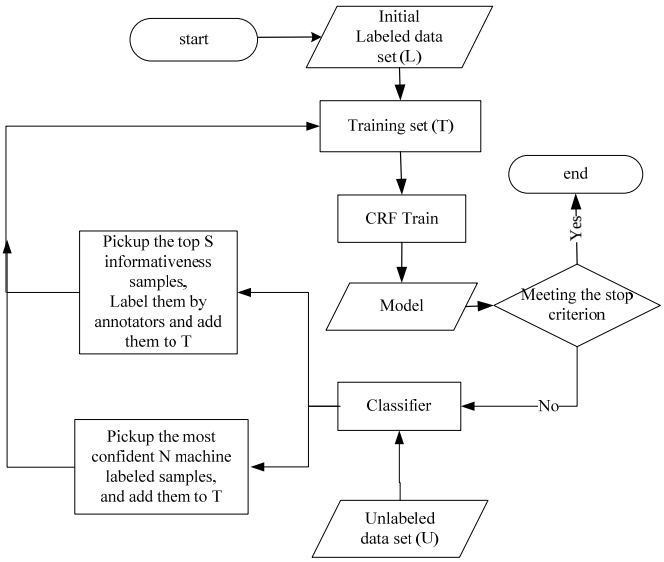

Figure 6. The combination of active learning and semi-supervised learning

V. SYSTEM DESCRIPTIONS

A. System Procedure

CRFs model is chosen as the learner in the proposed combining framework. The whole procedure is shown in the Fig. 6. In the beginning, a certain amount of sentences are randomly abstracted and labeled as initial training set for the CRF model. To increase the training set, new samples are picked up from the unlabeled data set based on information density strategy. The data with less information density is labeled by annotators and added into the training set. The machine-labeled data with top information density are also been input into the training set. The mixture of those data joins the retrain processing. The selecting and labeling procedures are iterated until stopping criterion is met.

B. Features

The feature used in our CRFs model includes two types: context features and dictionary features.

Context features include the characters prior and after the target character. The window size is 9.

In order to make use of human knowledge in NER task, dictionary features were involved in our system. 9 difference dictionary and gazetteer are considered as shown in Table I. The Table II shows the details of dictionary features.

The first column in Table II presents the character combinations. C_0 indicates the target character; C with subscript represents the nearby characters, negative value means the character is before the target character, positive value means that it is after the target character. The value of the subscript number shows relative position compared with current character. The C_0C_1 denotes the combination of the current character and it's immediate subsequence.

TABLE I. DICTIONARY DESCRIPTIONS

Dictionary Number	Dictionary Name
1	Chinese Surname Dictionary
2	General Chinese Name Characters Dictionary
3	Transliteration Characters Dictionary
4	Location Name Dictionary
5	Location Suffix Dictionary
6	Organization Suffix Dictionary
7	Single Character List
8	Title List
9	Chinese Name List

TABLE II. DICTIONARY FEATURE TEMPLATE

Character Combination	Checked Dictionary Numbers
C_0	1，2，3，4，5，6，7
$C_{-1}C_0$	8，4，5，6，9
$C_{-2}C_{-1}C_0$	4，6，8，9
$C_{-3}C_{-2}C_{-1}C_0$	4，5，6，8
$C_{-4}C_{-3}C_{-2}C_{-1}C_0$	4，5，6，8
C_0C_1	9
$C_0C_1C_2$	9
$C_0C_1C_2C_3$	9
$C_{-1}C_0C_1$	9
$C_{-2}C_{-1}C_0C_1$	9
$C_{-1}C_0C_1C_2$	9

For example, given a character sequence “在中国致公党第十一次全国代表大会隆重召开之际/ on the moment of holding the eleventh National Party Congress for China Zhi Gong Dang”, when considering the current character C_0 denotes “党”, C_{-4} denotes “中” C_{-3} denotes “国”, C_{-2} denotes “致”, C_{-1} denotes “公”. When preparing data, we match the training and testing data first by the dictionaries, and then get features from the matched data. According the feature template shown in the prior table, the sequence $C_{-4}C_{-3}C_{-2}C_{-1}C_0$ “中国致公党”shall be checked and this character sequence will be found out as a word listed in the dictionary 6. And this is only one feature about this sequence. Huge amount features are extracted during the training processing and finally the model is generated. During the procedure of decoding, the most possible tag for a particular sequence is found out.

VI. EXPERIMENTS

A. Data Sets

The training corpus used in our experiment is Sighan bakeoff 2006 MSRA corpus[1]. The detail of corpus is shown in Table III. Table IV presents the entity distribution in the training and test data.

TABLE III. DATA DESCRIPTIONS

Data Set	# of sentences	# of character
Training set	136,621	2,170,848
Test set	11,504	172,602

[1] http://www.sighan.org/bakeoff2006/

B. Active Learning Experiment

Before combining with the semi-supervised learning method, the active learning is used separately. The size of initial training set L is chosen as 100. For each training round, the top 100 most informative samples are picked up, labeled by annotators and added into the training set. The result for the process before 50 iterations is shown in Fig. X. Comparing it with random picking up, we can see that it is obvious that the active learning algorithm enhance the system effectively, especially when the data size is small. In our system, the higher growth rate is achieved before 20 iterations, which only included 2,000 labeled sentences.

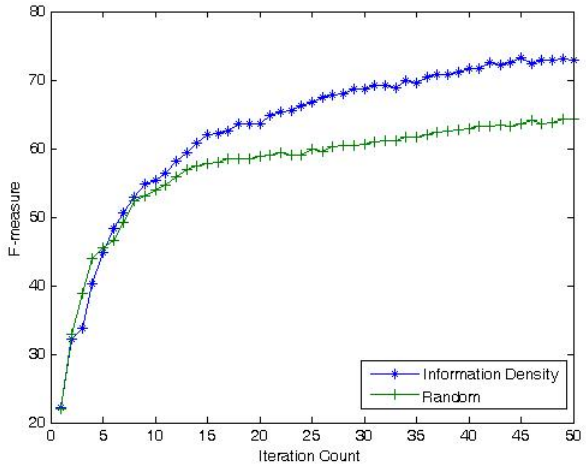

Figure 7. Experimental Result based on Active Learning Strategy

C. Experiment settings

In our experiment setting, to ensure the training quality of semi-supervised learning, the initial training set L is 5000. The iteration procedure is stopped until the improvement of performance tends to stabilize. The iteration count is set to 70. In order to calculate the proposed modified information density, the 20-best labeled results for each sentence were output by the CRFs model. In each iteration, the top 100 and bottom 50 samples according to the modified information density score were picked up from the unlabeled pool. The top 100 samples are considered as the most informative samples which embedding the knowledge that current model haven't known yet. They will be labeled by annotators and added into the training set. The bottom 50 samples are the most confidently labeled samples which could be used to reinforce the knowledge current model already got. They will be put into the training set with the labels given by the model. The performance for the whole process is shown in Fig. 7. We can observe from Fig. 7 that the combined method clearly enhances the system effectively, especially when the data size is small.

TABLE IV. ENTITY DISTRIBUTION IN TRAINING DATA AND TEST DATA

Data Set	# of person name	# of location name	# of organization name	total
Training set	17,615	36,860	20,581	75,056
Test set	1,973	2,886	1,333	6,190

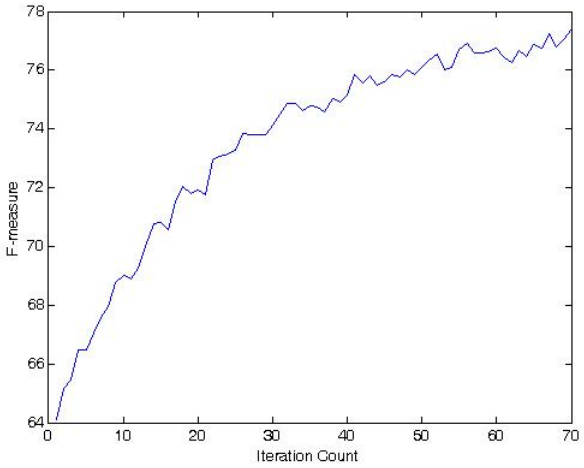

Figure 8. Experimental Result based on the hybrid method

In order to show the effective of proposed method, further we randomly pickup 10,000, 20,000 and 130,000 labeled samples as training set for the general CRF respectively, the test result are shown in Table V. The results in the first two rows of Table 5 are distinct lower than the results in fourth row, which are achieved using the approach of combing self-training and active learning. In spite of the result of combining method is lower than the result of the general CRF with 130,000 sentences, the labeled data requirement of the hybrid method is much smaller than usual. Since the initial size is different, the pure active learning result is not listed in this table.

TABLE V. EXPERIMENTAL RESULT

Training Data Size	P(%)	R(%)	F(%)
Random 10,000	81.3	63.8	71.5
Random 20,000	81.2	67.0	73.4
Total 130,000	85.0	76.7	80.7
(Avtive + Self-training) 15,500	83.75	72.02	77.4

VII. CONLUSIONS

We addressed Chinese Named Entity Recognition by the combination of pool-based active learning algorithm and semi-supervised learning methods in this paper. Self-training is selected as the method of semi-supervised learning. Conditional Random Fields model is chosen as the underlying model for the hybrid method. Rich dictionary features for the CRF model have been adopted. A modified Information Density is proposed as the criteria for sample selecting. Not only the average information similarities between target samples with unlabeled samples have been taken into account but also similarities with labeled samples set were considered in the proposed criteria. The new strategy we proposed makes the data with representative information have much higher selection opportunity and improve the system learning ability effectively.

The hybrid algorithm combining with conditional random field lessens the dependence on huge labeled training data, which is time consuming and expensive. In our system, only one tens labeled sentences are selected to achieve the similar performance got by general CRF with total labeled training samples.

In this work, fix amount training samples are added for each iteration, we will find how to use the unlabeled data chose by self training in a more intelligent way in future work, such as increasing the added samples with the increasement of performance. Also, we will try more underlying basic classifiers to improve the time efficiency of the proposed combing framework.

REFERENCE

[1] O. Bender, F. Och, and H. Ney, "Maximum entropy models for named entity recognition," Proceedings of CoNLL-2003, 2003, pp. 148-151.
[2] H. Chieu and H. Ng, "Named entity recognition with a maximum entropy approach," In Proceedings of CoNLL-2003, 2003.
[3] J. Curran and S. Clark, "Language independent NER using a maximum entropy tagger," Proceedings of the seventh conference on Natural language learning at HLT-NAACL , 2003, pp. 164-167.
[4] H. Chieu and H. Ng, "Named entity recognition: a maximum entropy approach using global information," Proceedings of COLING02, 2002, pp. 190–196.
[5] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," Proceedings of CoNLL-2003, 2003, pp. 168-171.
[6] D. Klein, J. Smarr, H. Nguyen, and C. Manning, "Named entity recognition with character-level models," Proceedings of CoNLL, 2003.
[7] J. Mayfield, P. McNamee, and C. Piatko, "Named entity recognition using hundreds of thousands of features," Proceedings of CoNLL , 2003, pp. 184-187.
[8] C. Whitelaw and J. Patrick, "Named entity recognition using a character-based probabilistic approach," Proceedings of CoNLL, 2003, pp. 196-199.
[9] T. Makino and Y. Ohta, "Tuning support vector machines for biomedical named entity recognition," In Proc. of ACL-02 Workshop on Natural Language Processing in the Biomedical Domain , 2002, pp. 1-8.
[10] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," Proceedings of CoNLL , 2003.
[11] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets", International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004, pp. 104–107.
[12] F. Jiao, S. Wang, C. Lee, R. Greiner, and D. Schuurmans, "Semi-supervised conditional random fields for improved sequence segmentation and labeling," 2006, p. 209.
[13] D. Angluin, "Queries and concept learning," *Machine learning*, vol. 2, pp. 319-342, 1988.

[14] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *In: Proc. Internat. Conf. on Machine Learning (ICML)*, Palo Alto, CA, 2000, pp. 839-846.

[15] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45-66, 2002.

[16] C. Symons, N. Samatova, R. Krishnamurthy, B. Park, T. Umar, D. Buttler, T. Critchlow, and D. Hysom, "Multi-criterion active learning in conditional random fields," 2006, pp. 323–331.

[17] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," 1998.

[18] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 3-12.

[19] A. McCallum, "Employing EM in pool-based active learning for text classification," in *Proceedings of the International Conference on Machine Learning (ICML)*, Morgan Kaufmann, 1998, pp. 359–367.

[20] S. Hoi, R. Jin, and M. Lyu, "Large-scale text categorization by batch mode active learning," in *Proceedings of the International Conference on the World Wide Web*, 2006a., pp. 633-642.

[21] C. Thompson, M. Califf, and R. Mooney, "Active learning for natural language parsing and information extraction," in *Proceedings of the International Conference on Machine Learning (ICML)*, Morgan Kaufmann, 1999, pp. 406-414.

[22] B. Settles and M. Craven., "An analysis of active learning strategies for sequence labeling tasks.," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008, pp. 1069–1078.

[23] M. Tang, X. Luo, and S. Roukos, "Active learning for statistical natural language parsing," in *Proceeding Internatinal Conference on Machine Learning (ICML)*, Bled, Sloveni, 2001, pp. 120-127.

[24] M. Sassano, "An empirical study of active learning with support vector machines for Japanese word segmentation," in *Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA., 2002, pp. 505-512.

[25] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," 2001, pp. 107-118.

[26] C. Zhang and T. Chen, "An active learning framework for content-based information retrieval," *IEEE transactions on multimedia*, vol. 4, pp. 260-268, 2002.

[27] R. Yan, J. Yang, and A. Hauptmann, "Automatically labeling video data using multi-class active learning," in *Proceedings of the International Conference on Computer Vision*, 2003, pp. 516–523.

[28] A. Hauptmann, W. Lin, R. Yan, J. Yang, and M. Chen, "Extreme video retrieval: joint maximization of human and computer performance," 2006, pp. 385-394.

[29] G. Tur and D. Hakkani-Tu¨ r, "Unsupervised learning for spoken language understanding," in *Proceeding of European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland., September 2003.

[30] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proceeding of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, FL, 2002, pp. 3904-3907.

[31] Y. Liu, "Active learning with support vector machine applied to gene expression data for cancer classification," *Journal of chemical information and computer sciences*, vol. 44, pp. 1936-1941, 2004.

[32] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," 1995, pp. 189-196.

[33] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," 2000, pp. 86-93.

[34] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," in *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*,2003, pp. 25–32.

[35] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Seventh IEEE Workshop on Applications of Computer Vision*, 2005, pp. 29–36.

[36] M. Collins and Y. Singer, "Unsupervised models for named entity classification," 1999, pp. 189–196.

[37] R. Jones, "Learning to extract entities from labeled and unlabeled text," Carnegie Mellon University., Doctoral Dissertation CMU-LTI-05-191, 2005.

[38] A. McCallum, "Employing EM in pool-based active learning for text classification," 1998.

[39] I. Muslea, S. Minton, and C. Knoblock, "Active+ semi-supervised learning= robust multi-view learning," in *Internat. Conf. on Machine Learning (ICML)*, Sydney, Australia., 2002, pp. 435-442.

[40] G. Riccardi and D. Hakkani-Tur, "Active and unsupervised learning for automatic speech recognition," 2003.

[41] L. Rabiner, "A tutorial on hidden Markov models and selected applications inspeech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.

[42] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proceeding 18th International Conference on Machine Learning 2001, pp. 282-289.

[43] H. Wallach, "Conditional Random Fields: An Introduction," *Rapport technique MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania*, vol. 50, 2004.

[44] X. Zhu, J. Lafferty, and R. Rosenfeld, "Semi-supervised learning with graphs," Doctoral dissertation CMU-LTI-05-192, School of Computer Science, Carnegie Mellon University, 2005.

[45] B. Settles, "Active Learning Literature Survey," In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2009.