Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

CRF-based Active Learning for Chinese Named Entity Recognition*

Lin Yao[1], Chengjie Sun[2], Shaofeng Li[1], Xiaolong Wang[1], Xuan Wang[1]

[1]Computer Science Department, HITSGS
ShenZhen, China
[2] School of Computer Science and Technology, Harbin Institute of Technology
Harbin, China
{yaolin, cjsun,Wangxl, wangxuan}@insun.hit.edu.cn, vicholi@cs.hitsz.edu.cn

***Abstract*—Conditional Random Fields (CRFs) have been used for many sequence labeling tasks and got excellent results. Further, the supervised model strongly depends on the huge training data. Active learning is a different way rather than relying on a large amount random sampling. However, random sampling constructively participates in the optimal choosing training examples. Based on different query strategies, active learning can combine with other machine learning methods to reduce the annotation cost while maintaining the accuracy. This paper proposes a new active learning strategy based on Information Density (ID) integrated with CRFs for Chinese Named Entity Recognition (NER). On Sighan bakeoff 2006 MSRA NER corpus, an F1 score of 77.2% is achieved by using only 10,000 labeled training sentences chosen by the proposed active learning strategy.**

***Keywords*—conditional random field, active learning, named entity recognition, information density**

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 and must rely on large labeled training data with high quality. However, the shortage of annotated training sets limits the further development, especially for some languages which are not well know. Another problem coming with huge data set is the large consumption of training time.

To reduce the number of labeled training samples to shorten the training time and to achieve the requested performance is not a trivial task. 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.

In this paper, we propose an alternative active learning strategy for the NER task. Without large-scale labeled data, the proposed method greatly reduces the training time 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 1, Section 2 presents an introduction of the general CRFs framework. We present the feature template used in our system in Section 3. Section 4 gives the description of active learning model, followed by our new query strategy based on information density. We present and analyze our results in section 5. Finally, some concluding remarks are presented in Section 6.

II. CONDITIONAL RANDOM FIELD

Similar to the Hidden Markov Models (HMMs)[12], Conditional Random Fields (CRFs)[13] 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.

A linear-chain CRFs model is described as following:

* 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)

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

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

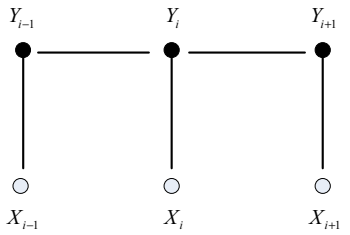

Figure 1. Graphical representation of liner-chain CRFs

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

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 [14]. Each feature f_k can be either state feature function or transition feature function. $\lambda's$ are the parameters to weight corresponding features and can be estimated from training data. Z is a normalization factor.

III. 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 [15] 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.

A. Active Learning Scenarios

Based on different active learning settings, there are 3 main types active learning [16]: (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
    θ = 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∈U} φ(x)
    // Add selected subset into training set L
    L = L ∪ 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.

B. 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.

C. 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).

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

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

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 $\left(\frac{1}{U}\sum_{u=1}^{U}Sim\left(x,x^{u}\right)\right)^{\beta}$ which is investigated based on Cosine similarity function. The similarity with labeled data is shown as $e^{-sim\left(x,x^{L}\right)}$.

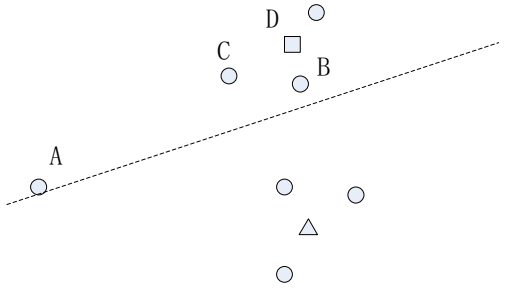

Figure 2. Analysis for uncertainty sampling

The sequence of feature vectors X, is represented by the matrix as shown in Figure 1. Each token in a sequence is described by a feature vector $\left(f_1 \ldots f_J\right)$; 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).

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

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 $\sum_{t=1}^{T}f_j\left(x_t\right)$ 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.

IV. SYSTEM DESCRIPTIONS

A. Active Learning Procedure

CRFs model is chosen as the learner in the proposed active learning framework. The whole procedure is shown in the Fig. 4. 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, labeled by annotators and added into the training set. The sampling and labeling procedures are iterated until stopping criterion is met.

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.

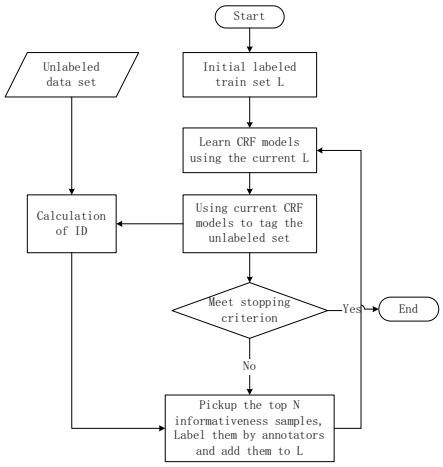

Figure 4. Active learning procedure

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.

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
$C_{-3}C_{-2}C_{-1}C_0$	4，5，6，8，9
$C_{-4}C_{-3}C_{-2}C_{-1}C_0$	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

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. For example, the C_0C_1 denotes the combination of the current character and it's immediate subsequence.

V. 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 III. DATA DESCRIPTIONS

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

B. Experiment settings

In our experiment setting, the size of initial training set L is 100. The iteration time for active learning is 100. In order to calculate the sequence entropy, the 20-best tag results for each sentence were output by the CRFs model. During each iteration, the top 100 most informative samples were picked up, labeled by annotators and added into the training set. The result for the whole process is shown in Fig. 5. We can observe from Fig. 5 that the active learning algorithm clearly enhances 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.

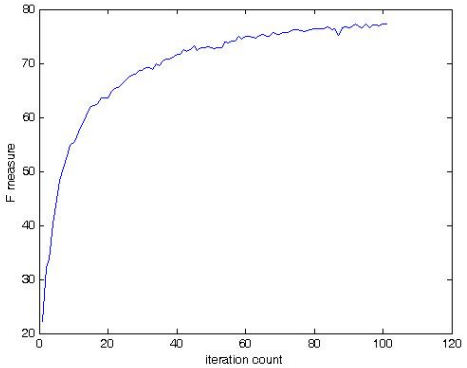

Figure 5. Experimental Result based on Active Learning Strategy

In order to show the effective of active learning, 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 IV. The results in the first two rows of Table 5 are distinct lower than the results in fourth row, which

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

are achieved using active learning algorithm. In spite of the result of active learning is lower than the result of the general CRF with 130,000 sentences, the labeled data requirement of active learning is much smaller than usual.

TABLE IV. 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
Active Learning 10,000	83.2	72.1	77.2

VI. CONLUSIONS

We addressed in this paper Chinese Named Entity Recognition by pool-based active learning algorithm based on Conditional Random Field. A rich dictionary features for the CRF model have been adopted. The pool-based active learning algorithm used in our system was based on Information Density. 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. The proposed new strategy makes the data with representative information have much higher selection opportunity and improve the system learning ability effectively.

The active learning algorithm combining with conditional random field lessens the dependence on huge labeled training data, which is time consuming and expensive. In our system, only 10,000 labeled sentences are selected to achieve the similar result got by general CRF with 130,000 labeled training samples.

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

[13] 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.

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

[15] 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.

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