On Using Ensemble Methods for Chinese Named Entity Recognition

Chia-Wei Wu **Shyh-Yi Jan** **Tzong-Han Tsai** **Wen-Lian Hsu**

Institute of Information Science, Academia Sinica, Nankang, Taipei,115, Taiwan
{cwwu,shihyi,thtsai,Hsu}@iis.sinica.edu.tw

Abstract

In sequence labeling tasks, applying different machine learning models and feature sets usually leads to different results. In this paper, we exploit two ensemble methods in order to integrate multiple results generated under different conditions. One method is based on majority vote, while the other is a memory-based approach that integrates maximum entropy and conditional random field classifiers. Our results indicate that the memory-based method can outperform the individual classifiers, but the majority vote method cannot.

1 Introduction

Sequence labeling and segmentation tasks have been studied extensively in the fields of computational linguistics and information extraction. Several tasks, including, word segmentation, and semantic role labeling, provide rich information for various applications, such as segmentation in Chinese information retrieval and named entity recognition in biomedical literature mining.

Probabilistic state automata models, such as the Hidden Markov model (HMM) [6] and conditional random fields (CRF) [5] are some of best, and therefore most popular, approaches for sequence labeling tasks. Both HMM and CRF consider that the state transition and the state prediction are conditional on the observation of data. The advantage of the CRF model is that richer feature sets can be considered, because, unlike HMM, it does not make a dependence assumption. However, the obvious drawback of the CRF model is that it needs more computing resources, so we can not apply all the features of the model. One possible way to resolve this problem is to effectively combine the results of various individual classifiers trained with different feature sets. In this paper, we use two ensemble methods to combine the results of the classifiers. We also combine the results generated by two machine learning models: maximum entropy (ME) [1] and CRF. One ensemble method is based on the majority vote [3], and the other is the memory based learner [7]. Although the ensemble methods have been applied in some sequence labeling tasks [2],[3], similar work in Chinese named entity recognition is scarce.

Our Chinese named entity tagger uses a character-based model. For English named entity tasks, a character-based NER model proposed by Dan Klein [4] proves the usefulness of substrings within words. In Chinese NER, the character-based model is more straightforward, since there are no spaces between Chinese words and each Chinese character is actually meaningful. Another reason for using a character-based model is that it can avoid the errors sometimes made by a Chinese word segmentor.

The remainder of this paper is organized as follows. In the Section 2, we introduce the machine learning models, the features we apply in the machine learning models, and the ensemble methods. In Section 3, we briefly describe the experimental data and the experiment results. Then, in Section 4, we present our conclusions..

2 Method

2.1 Machine Learning Models

In this section, we introduce ME and CRF.

Maximum Entropy

ME[1] is a statistical modeling technique used for estimating the conditional probability of a target label based on given information. The technique computes the probability $p(y|x)$, where y denotes all possible outcomes of the space, and x denotes all possible features of the space. The computation of $p(y|x)$ depends on a set of features in x; the features are helpful for making predictions about the outcomes, y.

Given a set of features and a training set, the ME estimation process produces a model, in which every feature f_i has a weight λ_i. The ME model can be represented by the following formula:

$$p(y \mid x) = \frac{1}{z(x)} \exp\left(\sum_i \lambda_i f_i(x,y) \right),$$

$$z(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y) \right).$$

The probability is derived by multiplying the weights of the active features (i.e., those f_i (y,x) = 1).

Conditional Random Field

A conditional random field (CRF)[5] can be seen as an undirected graph model in which the nodes corresponding to the label sequence ***y*** are conditional on the observed sequence ***x***. The goal of CRF is to find the label sequence ***y*** that has the maximized probability, given an observation sequence ***x***. The formula for the CRF model can be written as:

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_j \lambda_j F_j(\mathbf{y}, \mathbf{x}) \right),$$

where λ_j is the parameter of a corresponding feature F_j , Z(x) is an normalizing factor, and F_j can be written as:

$$F_j(\mathbf{y}, \mathbf{x}) = \sum_{i=0}^{n} f_i(y_{i-1}, y_i, \mathbf{x}, i),$$

where *i* means the relative position in the sequence, and y_{i-1} and y_i denote the label at position *i-1* and *i* respectively. In this paper, we only consider linear chain and first-order Markov assumption CRFs. In NER applications, a feature function f_j (y_{i-1}, y_i, x, i) can be set to check whether x is a specific character, and whether y_{i-1} is a label (such as *Location*) and y_i is a label (such as *Others*).

2.2 Chinese Named Entity Recognition

In this section, we present the features applied in our CRF and ME models, namely, characters, words, and chuck information.

Character Features

The character features we apply in the CRF model and the ME model are presented in Tables 1 and 2 respectively. The numbers listed in the feature type column indicate the relative position of a character in the sliding window. For example, -1 means the previous character of the target character. Therefore, the characters in those positions are applied in the model. The numbers in parentheses mean that the feature includes a combination of the characters in those positions. The unigrams in Tables 1 and 2 indicate that the listed features only consider to their own labels, whereas the bigram model considers the combination of the current label and the previous label. Since ME does not consider multiple states in a single feature, there are only unigrams in Table 2. In addition, as ME can handle more features than CRF, we apply extra features in the ME model

Table 1 Character features for CRF

	Feature Types
unigram	-2, -1, 0, 1, 2, (-2,-1), (-1,0), (0,1), (1,2), (-1,0,1)
bigram	-2 -1 0 +1 +2, (0,1)

Table 2 Character features for ME

	Feature Types
unigram	-2, -1, 0, 1, 2, (-2,-1), (-1,0), (0,1), (1,2), (-1,0,1) (-1,1)

Word Information

Because of the limitations of the closed task, we use the NER corpus to train the segmentors based on the CRF model. To simulate noisy word information in the test corpus, we use a ten-fold method for training segmentors to tag the training corpus. The word features we apply in our NER systems are presented in Tables 3 and 4.

In addition to the word itself, chuck information, i.e., the relative position of a character in a word, is also valuable information. Hence, we also add chuck information to our models. As the diversity of Chinese words is greater than that of Chinese characters, the number of features that can be used in CRF is much lower than the number that can be used in ME.

Table 3 Word features for CRF

	Feature Types
unigram	0
bigram	0

Table 4 Word features for ME

	Feature Types
unigram	-1, 0, 1, (-2,-1), (-1,0), (0,1), (1,2)

2.3 Ensemble Methods

Majority vote

We can not put all the features into the CRF model because of its limited resources. Therefore, we train several CRF classifiers with different feature sets so that we can use as many features as possible. Then, we use the following simple,

equally weighted linear equation, called majority vote, to combine the results of the CRF classifiers.

$$S(y,x) = \sum_{i=0}^{T} C_i(y,x),$$

where S(y,x) is the score of a label y and a character x respectively; T denotes the total number of CRF models; and the value of C_i(y,x) is 1 if the decision of the result of the i_{th} CRF model is y, otherwise it is zero. The highest score of y is chosen as the label of x. The results are incorporated into the Viterbi algorithm to search for the path with the maximum scores.

In this paper, the first step in the majority vote experiment is to train three CRF classifiers with different feature sets. Then, in the second step, we use the results obtained in the first step to generate the voting scores for the Viterbi algorithm.

Memory Based learner

The memory-based learning method memorizes all examples in a training corpus. If a word is unknown, the memory-based classifier uses the k-nearest neighbors to find the most similar example as the answer. Instead of using the complete algorithm of the memory-based learner, we do not handle unseen data. In our memory- based combination method, the learner remembers all named entities from the results of the various classifiers and then tags the characters that were originally tagged as "Other". For example, if a character x is tagged by one classifier as "0" ("Others" tag) and if the memory-based classifier learns from another classifier that this character is tagged as PER, then x will be tagged as "B-PER" by the memory-based classifier.

The obvious drawback of this method is that the precision rate might decrease as the recall rate increases. Therefore, we set the following three rules to filter out samples that are likely to have a high error rate.

1. Named entities can not be tagged as different named entity tags by different classifiers.
2. We set an absolute frequency threshold to filter out examples that occur less than the threshold.
3. We set a relative frequency threshold to filter out examples that occur less than the threshold. For example, if a word x appears 10 times in the corpus, then half of the instances of x have to be tagged as named entities; otherwise, x will be filtered out of the memory classifier.

In our experiment, we used the memory-based learner to memorize the named entities from the tagging results of an ME classifier and a CRF classifier, and then tagged the tagging results of the CRF classifier.

3 Experiments

3.1 Data

We selected the corpora of City University of Hong Kong (CityU) and Microsoft Research (MSRA) corpora to evaluate our methods. CityU is a Traditional Chinese corpus, and MSRA is Simplified Chinese corpus.

3.2 Results

Table 5 shows the results of several methods applied to the MSRA corpus. The memory-based ensemble method, which combines the results of a maximum entropy model and those of a CRF classifier, achieves the best performance. The majority vote combined with the results of three CRF models based on different feature sets has the worst performance.

Table 5 msra

	Precision	Recall	FB1
Memory based	86.21	78.14	81.98
Majority Vote	85.83	76.06	80.65
Only-Character	86.70	75.54	80.74
CRF	86.23	77.40	81.58

The results obtained on Cityu, presented in Table 6, show that the single CRF classifier achieved the best performance. None of the ensemble methods can outperform the non-ensemble methods.

Table 6 cityu

	Precision	Recall	FB1
Memory based	90.79	86.26	88.47
Majority Vote	90.52	84.15	87.22
Only-Character	91.32	84.55	87.80
CRF	92.01	85.45	88.61

Tables 7 and 8 show the results of the memory-based ensemble methods under different rules. We set the frequency threshold as 2 and the relative frequency threshold as 0.5. The results show that the relative frequencies rule effectively reduces the loss of precision caused by more entities being tagged by the memory-based classifier. The memory-based ensemble method works well on the MSRA corpus, but not on the CityU corpus. In the MSRA corpus, the memory-based

ensemble method outperforms the individual CRF model by approximately 0.4 % in FB1. We found that the memory-based classifier can not achieve a better performance than the CRF model because it misclassifies many organizations' names. Therefore, we chose another strategy that restricts the memory-based classifier to tagging person names only. Under this restriction, the performance of the memory-based classifier improves FB1 by approximately 0.2%.

Table 7 msra- The performances of memory based ensemble methods under different rules.

	Precision	Recall	FB1
Frequency Threshold	86.18	78.16	81.97
Relative Frequency Threshold	86.21	78.14	81.98
Only Person	86.27	77.58	81.69

Table 8 cityu- The performances of memory based ensemble methods under different rules.

	Precision	Recall	FB1
Frequency Threshold	90.69	86.55	88.57
Relative Frequency Threshold	90.87	86.29	88.52
Only Person	92.00	85.66	88.72

4 Conclusion

In this paper, we use ME and CRF models to train a Chinese named entity tagger. Like previous researchers, we found that CRF models outperform ME models. We also apply two ensemble methods, namely, majority vote and memory-based approaches, to the closed NER shared task. Our results show that integrating individual classifiers as the majority vote approach does not outperform the individual classifiers. Furthermore, a memory-based combination only seems to work when we restrict the memory-based classifier to handling person names.

Acknowledgement

We are grateful for the support of National Science Council under Grant NSC 95-2752-E-001-001-PAE.

References

1. Berger, A., Pietra, S.A.D. and Pietra, V.J.D. A Maximum Entropy Approach to Natural Language Processing. Computer Linguistic, 22. 1996 39-71.
2. Florian, R., Ittycheriah, A., Jing, H. and Zhang, T., Named Entity Recognition through Classifier Combination. in Proceedings of Conference on Computational Natural Language Learning, 2003, 168-171.
3. Halteren, H.v., Zavrel, J. and Daelemans, W. Improving accuracy in word class tagging through combination of machine learning systems. Computational Linguistics, 27 (2). 2001 199-230.
4. Klein, D., Smarr, J., Nguyen, H. and Manning, C.D., Named Entity Recognition with Character-Level Models. in Conference on Computational Natural Language Learning, 2003, 180-183.
5. Lafferty, J., McCallum, A. and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. International Conference on Machine Learning. 2001 282-289.
6. Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77 (2). 1989 257-286.
7. Sutton, C., Rohanimanesh, K. and McCallum, A., Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. in Proceedings of the Twenty-First International Conference on Machine Learning, 2004, 99-107.
8. Zavrel, J. and Daelemans, W. Memory-based learning: using similarity for smoothing. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. 1997 436 - 443.