Selecting Optimal Feature Template Subset for CRFs

Xingjun Xu[1] and **Guanglu Sun**[2] and **Yi Guan**[1] and
Xishuang Dong[1] and **Sheng Li**[1]
1: School of Computer Science and Technology,
Harbin Institute of Technology,
150001, Harbin, China
2: School of Computer Science and Technology,
Harbin University of Science and Technology
150080, Harbin, China
xxjroom@163.com; guanglu.sun@gmail.com
guanyi@hit.edu.cn; dongxishuang@gmail.com
lisheng@hit.edu.cn

Abstract

Conditional Random Fields (CRFs) are the state-of-the-art models for sequential labeling problems. A critical step is to select optimal feature template subset before employing CRFs, which is a tedious task. To improve the efficiency of this step, we propose a new method that adopts the maximum entropy (ME) model and maximum entropy Markov models (MEMMs) instead of CRFs considering the homology between ME, MEMMs, and CRFs. Moreover, empirical studies on the efficiency and effectiveness of the method are conducted in the field of Chinese text chunking, whose performance is ranked the first place in task two of CIPS-ParsEval-2009.

1 Introduction

Conditional Random Fields (CRFs) are the state-of-the-art models for sequential labeling problem. In natural language processing, two aspects of CRFs have been investigated sufficiently: one is to apply it to new tasks, such as named entity recognition (McCallum and Li, 2003; Li and McCallum, 2003; Settles, 2004), part-of-speech tagging (Lafferty et al., 2001), shallow parsing (Sha and Pereira, 2003), and language modeling (Roark et al., 2004); the other is to exploit new training methods for CRFs, such as improved iterative scaling (Lafferty et al., 2001), L-BFGS (McCallum, 2003) and gradient tree boosting (Dietterich et al., 2004).

One of the critical steps is to select optimal feature subset before employing CRFs. McCallum (2003) suggested an efficient method of feature induction by iteratively increasing conditional log-likelihood for discrete features. However, since there are millions of features and feature selection is an NP problem, this is intractable when searching optimal feature subset. Therefore, it is necessary that selects feature at feature template level, which reduces input scale from millions of features to tens or hundreds of candidate templates.

In this paper, we propose a new method that adopts ME and MEMMs instead of CRFs to improve the efficiency of selecting optimal feature template subset considering the homology between ME, MEMMs, and CRFs, which reduces the training time from hours to minutes without loss of performance.

The rest of this paper is organized as follows. Section 2 presents an overview of previous work for feature template selection. We propose our optimal method for feature template selection in Section 3. Section 4 presents our experiments and results. Finally, we end this paper with some concluding remarks.

2 Related Work

Feature selection can be carried out from two levels: feature level (feature selection, or FS), or feature template level (feature template selection, or FTS). FS has been sufficiently investigated and

share most concepts with FTS. For example, the target of FS is to select a subset from original feature set, whose optimality is measured by an evaluation criterion (Liu and Yu, 2005). Similarly, the target of FTS is to select a subset from original feature template set. To achieve optimal feature subset, two problems in original set must be eliminated: irrelevance and redundancy (Yu and Liu, 2004). The only difference between FS and FTS is that the number of elements in feature template set is much less than that in feature set.

Liu and Yu (2005) classified FS models into three categories: the filter model, the wrapper model, and the hybrid model. The filter model (Hall 2000; Liu and Setiono, 1996; Yu and Liu, 2004) relies on general characteristics of the data to evaluate and select feature subsets without any machine learning model. The wrapper model (Dy and Brodley, 2000; Kim et al., 2000; Kohavi and John, 1997) requires one predetermined machine learning model and uses its performance as the evaluation criterion. The hybrid model (Das, 2001) attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages.

There are two reasons to employ the wrapper model to accomplish FTS: (1) The wrapper model tends to achieve better effectiveness than that of the filter model with respect of a more direct evaluation criterion; (2) The computational cost is tractable because it can reduce the number of subsets sharply by heuristic algorithm according to the human knowledge. And our method belongs to this type.

Lafferty (2001) noticed the homology between MEMMs and CRFs, and chose optimal MEMMs parameter vector as a starting point for training the corresponding CRFs. And the training process of CRFs converges faster than that with all zero parameter vectors.

On the other hand, the general framework that processes sequential labeling with CRFs has also been investigated well, which can be described as follows:

1. Converting the new problem to sequential labeling problem;
2. Selecting optimal feature template subset for CRFs;
3. Parameter estimation for CRFs;
4. Inference for new data.

In the field of English text chunking (Sha and Pereira, 2003), the step 1, 3, and 4 have been studied sufficiently, whereas the step 2, how to select optimal feature template subset efficiently, will be the main topic of this paper.

3 Feature Template Selection

3.1 The Wrapper Model for FTS

The framework of FTS based on the wrapper model for CRFs can be described as:

1. Generating the new feature template subset;
2. Training a CRFs model;
3. Updating optimal feature template subset if the new subset is better;
4. Repeating step 1, 2, 3 until there are no new feature template subsets.

Let N denote the number of feature templates, the number of non-empty feature template subsets will be (2^N-1). And the wrapper model is unable to deal with such case without heuristic methods, which contains:

1. Atomic feature templates are firstly added to feature template subset, which is carried out by: Given the position i, the current word W_i and the current part-of-speech P_i are firstly added to current feature template subset, and then W_{i-1} and P_{i-1}, or W_{i+1} and P_{i+1}, and so on, until the effectiveness is of no improvement. Taking the Chinese text chunking as example, optimal atomic feature template subset is $\{W_{i-3}\sim W_{i+3}, P_{i-3}\sim P_{i+3}\}$;

2. Adding combined feature templates properly to feature template set will be helpful to improve the performance, however, too many combined feature templates will result in severe data sparseness problem. Therefore, we present three restrictions for combined feature templates: (1) A combined feature template that contains more than three atomic templates are not allowable; (2) If a combined feature template contains three atomic feature template, it can only contain at most one atomic word template; (3) In a combined template, at most one word is allowable between the two most adjacent atomic templates; For example, the combined feature templates, such as $\{P_{i-1}, P_i, P_{i+1}, P_{i+2}\}$, $\{W_i, W_{i+1}, P_i\}$, and $\{P_{i-1}, P_{i+2}\}$, are not allowable, whereas the combined templates, such as $\{P_i, P_{i+1}, P_{i+2}\}$, $\{P_{i-1}, W_i, P_{i+1}\}$, and $\{P_{i-1}, P_{i+1}\}$, are allowable.

3. After atomic templates have been added, $\{W_{i-1}, W_i\}$, or $\{W_i, W_{i+1}\}$, or $\{P_{i-1}, P_i\}$, or $\{P_i, P_{i+1}\}$ are firstly added to feature template subset. The template window is moved forward, and then backward. Such process will repeat with expanding template window, until the effectiveness is of no improvement.

Tens or hundreds of training processes are still needed even if the heuristic method is introduced. People usually employ CRFs model to estimate the effectiveness of template subset However, this is more tedious than that we use ME or MEMMs instead. The idea behind this lie in three aspects: first, in one iteration, the Forward-Backward Algorithm adopted in CRFs training is time-consuming; second, CRFs need more iterations than that of ME or MEMMs to converge because of larger parameter space; third, ME, MEMMs, and CRFs, are of the same type (log-linear models) and based on the same principle, as will be discussed in detail as follows.

3.2 Homology of ME, MEMMs and CRFs

ME, MEMMs, and CRFs are all based on the Principle of Maximum Entropy (Jaynes, 1957). The mathematical expression for ME model is as formula (1):

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{m} \lambda_i f_i(x, y)) \qquad (1)$$

, and Z(x) is the normalization factor.

MEMMs can be considered as a sequential extension to the ME model. In MEMMs, the HMM transition and observation functions are replaced by a single function $P(Y_i|Y_{i-1}, X_i)$. There are three kinds of implementations of MEMMs (McCallum et al., 2000) in which we realized the second type for its abundant expressiveness. In implementation two, which is denoted as MEMMs_2 in this paper, a distributed representation for the previous state Y_{i-1} is taken as a collection of features with weights set by maximum entropy, just as we have done for the observations X_i. However, label bias problem (Lafferty et al., 2001) exists in MEMMs, since it makes a local normalization of random field models. CRFs overcome the label bias problem by global normalization.

Considering the homology between CRFs and MEMMs_2 (or ME), it is reasonable to suppose that a useful template for MEMMs_2 (or ME) is also useful for CRFs, and vice versa. And this is a necessary condition to replace CRFs with ME or MEMMs for FTS.

3.3 A New Framework for FTS

Besides the homology of these models, the other necessary condition to replace CRFs with ME or MEMMs for FTS is that all kinds of feature templates in CRFs can also be expressed by ME or MEMMs. There are two kinds of feature templates for CRFs: one is related to Y_{i-1}, which is denoted as $g(Y_{i-1}, Y_i, X_i)$; the other is not related to Y_{i-1}, which is denoted as $f(Y_i, X_i)$. Both of them can be expressed by MEMMs_2. If there is only the second kind of feature templates in the subset, it can also be expressed by ME. For example, the feature function $f(Y_i, P_i)$ in CRFs can be expressed by feature template $\{P_i\}$ in MEMMs_2 or ME; and $g(Y_{i-1}, Y_i, P_i)$ can be expressed by feature template $\{Y_{i-1}, P_i\}$ in MEMM_2.

Therefore, MEMMs_2 or ME can be employed to replace CRFs as machine learning model for improving the efficiency of FTS.

Then the new framework for FTS will be:

1. Generating the new feature template subset;
2. Training an MEMMs_2 or ME model;
3. Updating optimal feature template subset if the new subset is better;
4. Repeating step 1, 2, 3 until there are no new feature template subsets.

The wrapper model evaluates the effectiveness of feature template subset by evaluating the model on testing data. However, there is a serious efficiency problem when decoding a sequence by MEMMs_2. Given N as the length of a sentence, C as the number of candidate labels, the time complexity based on MEMMs_2 is $O(NC^2)$ when decoding by viterbi algorithm. Considering the C different Y_{i-1} for every word in a sentence, we need compute $P(Y_i|Y_{i-1}, X_i)$ (N.C) times for MEMMs_2.

Reducing the average number of candidate label C can help to improve the decoding efficiency. And in most cases, the Y_{i-1} in $P(Y_i|Y_{i-1}, X_i)$ is not necessary (Koeling, 2000; Osbome, 2000). Therefore, to reduce the average number of candidate labels C, it is reasonable to use an ME model to filter the candidate label. Given a threshold T ($0 <= T <= 1$), the candidate label filtering algorithm is as follows:

1. CP = 0;
2. While CP <= T
   a) Add the most probable candidate label Y' to viterbi algorithm;
   b) Delete Y' from the candidate label set;
   c) $CP = P(Y'|X_i) + CP$.

If the probability of the most probable candidate label has surpassed T, other labels are discarded. Otherwise, more labels need be added to viterbi algorithm.

4 Evaluation and Result

4.1 Evaluation

We evaluate the effectiveness and efficiency of the new framework by the data set in the task two of

CIPS-ParsEval-2009 (Zhou and Li, 2010). The effectiveness is supported by high F-1 measure in the task two of CIPS-ParsEval-2009 (see Figure 1), which shows that optimal feature template subset driven by ME or MEMMs is also optimal for CRFs. The efficiency is shown by significant decline in training time (see Figure 3), where the baseline is CRFs, and comparative methods are ME or MEMMs.

We design six subsets of feature template set and six experiments to show the effectiveness and efficiency of the new framework. As shown in Table 1 and Table 2, the 1~3 experiments shows the influence of the feature templates, which are unrelated to Y_{i-1}, for both ME and CRFs. And the 4~6 experiments show the influence of the feature templates, which are related to Y_{i-1}, for both MEMMs_2 and CRFs. In table 1, six template subsets can be divided into two sets by relevance of previous label: 1, 2, 3 and 4, 5, 6. Moreover, the first set can be divided into 1, 2, and 3 by distances between features with headwords; the second set can be divided into 4, 5 and 6 by relevance of observed value. In order to ensure the objectivity of comparative experiments, candidate label filtering algorithm is not adopted.

Task2			
Rank	No.	boundary+type	boundary+type+relation
1	01	93.20	92.10
2	15	92.85	91.76
3	12	92.36	
4	10_a	92.11	90.94
5	10_b	92.11	90.94
6	17	91.98	89.85
7	10_c	91.76	90.63
8	10_d	91.75	90.67
9	14	91.29	90.13
10	00	90.39	88.88

Figure 1: the result in the task two of CIPS-ParsEval-2009

1	W_i, W_{i-1}, W_{i-2}, W_{i+1}, W_{i+2}, P_i, P_{i-1}, P_{i-2}, P_{i+1}, P_{i+2}, $W_{i-1}\_W_i$, $W_i\_W_{i+1}$, $W_{i-1}\_W_{i+1}$, $P_{i-1}\_P_i$, $P_{i-2}\_P_{i-1}$, $P_i\_P_{i+1}$, $P_{i-1}\_P_{i+1}$, $P_{i-1}\_P_i\_P_{i+1}$, $P_{i-2}\_P_{i-1}\_P_i$, $P_i\_P_{i+1}\_P_{i+2}$, $W_i\_P_{i+1}$, $W_i\_P_{i+2}$, $P_i\_W_{i-1}$, $W_{i-2}\_P_{i-1}\_P_i$, $P_i\_W_{i+1}\_P_{i+1}$, $P_{i-1}\_W_i\_P_i$, $P_i\_W_{i+1}$
2	W_{i-3}, W_{i+3}, P_{i-3}, P_{i+3}, $W_{i-3}\_W_{i-2}$, $W_{i+2}\_W_{i+3}$, $P_{i-3}\_P_{i-2}$, $P_{i+2}\_P_{i+3}$
3	W_{i-4}, W_{i+4}, P_{i-4}, P_{i+4}, $W_{i-4}\_W_{i-3}$, $W_{i+3}\_W_{i+4}$, $P_{i-4}\_P_{i-3}$, $P_{i+3}\_P_{i+4}$
4	Y_{i-1}
5	$Y_{i-1}\_P_i\_P_{i+1}$, $Y_{i-1}\_P_i$, $Y_{i-1}\_P_{i-1}\_P_i$
6	$Y_{i-1}\_P_{i-4}$, $Y_{i-1}\_P_{i+4}$

Table 1: six subsets of feature template set

id	Model	FT subset
1	ME vs. CRFs	1
2	ME vs. CRFs	1, 2
3	ME vs. CRFs	1, 2, 3
4	MEMMs vs. CRFs	1, 2, 4
5	MEMMs vs. CRFs	1, 2, 4, 5
6	MEMMs vs. CRFs	1, 2, 4, 5, 6

Table 2: six experiments

4.2 Empirical Results

The F-measure curve is shown in Figure 2. For the same and optimal feature template subset, the F-1 measure of CRFs is superior to that of ME because of global normalization; and it is superior to that of MEMMs since it overcomes the label bias.

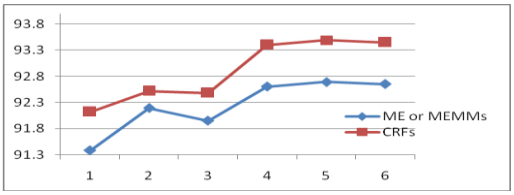

Figure 2: the F-measure curve

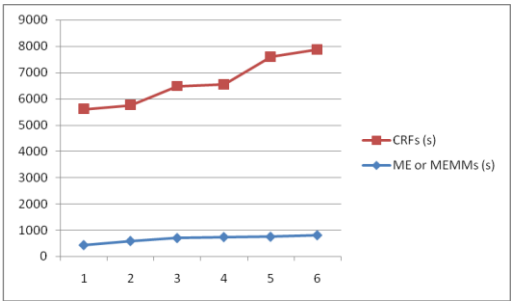

Figure 3: the training time curve

The significant decline in training time of the new framework is shown in Figure 3, while the testing time curve in Figure 4 and the total time curve in Figure 5. The testing time of ME is more

than that of CRFs because of local normalization; and the testing time of MEMMs_2 is much more than that of CRFs because of N.C times of $P(Y_i|Y_{i-1}, X_i)$ computation.

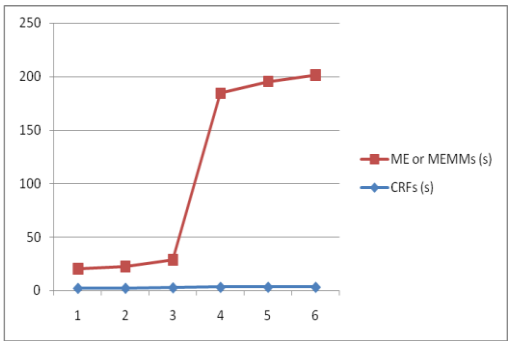

Figure 4: the testing time curve

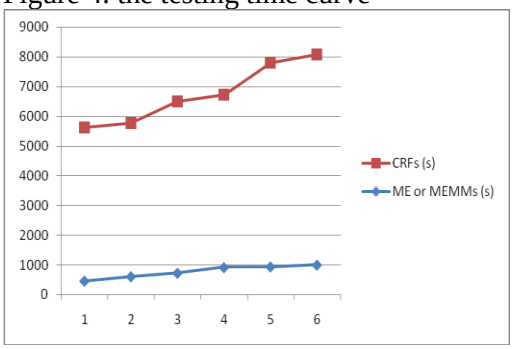

Figure 5: the total time curve

All results of ME and MEMMs in figures are represented by the same line because performances of these two models are the same when features are only related to observed values.

5 Conclusions

In this paper, we propose a new optimal feature template selection method for CRFs, which is carried out by replacing the CRFs with MEMM_2 (ME) as the machine learning model to address the efficiency problem according to the homology of these models. Heuristic method and candidate label filtering algorithm, which can improve the efficiency of FTS further, are also introduced. The effectiveness and efficiency of the new method is confirmed by the experiments on Chinese text chunking.

Two problems deserve further study: one is to prove the homology of ME, MEMMs, and CRFs theoretically; the other is to expand the method to other fields.

For any statistical machine learning model, feature selection or feature template selection is a computation-intensive step. This work can be adequately reduced by means of analyzing the homology between models and using the model with less computation amount. Our research proves to be a successful attempt.

References

Das Sanmay. 2001. Filters, wrappers and a boosting-based hybrid for feature selection. In Proceedings of the Eighteenth International Conference on Machine Learning, pages 74–81.

Dietterich Thomas G., Adam Ashenfelter, Yaroslav Bulatov. 2004. Training Conditional Random Fields via Gradient Tree Boosting. In Proc. of the 21th International Conference on Machine Learning (ICML).

Dy Jennifer G., and Carla E. Brodley. 2000. Feature subset selection and order identification for unsupervised learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 247–254.

Hall Mark A.. 2000. Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of the Seventeenth International Conference on Machine Learning, pages 359–366.

Jaynes, Edwin T.. 1957. Information Theory and Statistical Mechanics. Physical Review 106(1957), May. No.4, pp. 620-630.

Kim YongSeog, W. Nick Street and Filippo Menczer. 2000. Feature Selection in Unsupervised Learning via Evolutionary Search. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 365–369.

Koeling Rob. 2000. Chunking with Maximum Entropy Models. In Proceeding of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000, pp. 139-141.

Kohavi Ron, and George H. John. 1997. Wrappers for feature subset selection. Artificial Intelligence, 97(1-2):273–324.

Lafferty John, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning.

Li Wei, and Andrew McCallum. 2003. Rapid Development of Hindi Named Entity Recognition using Conditional Random Fields and Feature Induction. ACM Transactions on Asian Language Information Processing (TALIP).

Liu Huan, and Lei Yu. 2005. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on knowledge and Data Engineering, v.17 n.4, p.491-502.

Liu Huan, and Rudy Setiono. 1996. A probabilistic approach to feature selection - a filter solution. In Pro-

ceedings of the Thirteenth International Conference on Machine Learning, pages 319–327.

McCallum Andrew. 2003. Efficiently Inducing Features of Conditional Random Fields. In Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence.

McCallum Andrew, DAyne Freitag, Fernando Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In Proceedings of ICML'2000, Stanford, CA, USA, 2000, pp. 591-598.

McCallum Andrew, and Wei Li. 2003. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In Proceedings of The Seventh Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada.

Osbome Miles. 2000. Shallow Parsing as Part-of-speech Tagging. In Proceeding of CoNLL-2000 and LLL-2000, Lisbon, Portugal, 2000,pp. 145-147.

Roark Brian, Murat Saraclar, Michael Collins, and Mark Johnson. 2004. Discriminative language modeling with conditional random fields and the perceptron algorithm. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.

Settles Burr. 2004. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA).

Sha Fei, and Fernando Pereira. 2003. Shallow Parsing with Conditional Random Fields. Proceedings of the 2003 conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada.

Yu Lei, and Huan Liu. 2004. Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proceedings of the twentieth International Conference on Machine Learning, pages 856–863.

Zhou Qiang, and Yumei Li. 2010. Chinese Chunk Parsing Evaluation Tasks. Journal of Chinese Information Processing.