Applying Conditional Random Fields to Chinese Shallow Parsing

Yongmei Tan, Tianshun Yao, Qing Chen, and Jingbo Zhu

Natural Language Processing Lab,
Northeastern University, Shenyang 110004
yongmeitan@hotmail.com, chenqing@ics.neu.edu.cn
{tsyao, zhujingbo}@mail.neu.edu.cn

Abstract. Chinese shallow parsing is a difficult, important and widely-studied sequence modeling problem. CRFs are new discriminative sequential models which may incorporate many rich features. This paper shows how conditional random fields (CRFs) can be efficiently applied to Chinese shallow parsing. We employ using CRFs and HMMs on a same data set. Our results confirm that CRFs improve the performance upon HMMs. Our approach yields the F1 score of 90.38% in Chinese shallow parsing with the UPenn Chinese Treebank. CRFs have shown to perform well for Chinese shallow parsing due to their ability to capture arbitrary, overlapping features of the input in a Markov model.

1 Introduction

Chinese shallow parsing is an important component of most text analysis systems in applications such as information extraction and summary generation. This problem has been widely studied and approached from different aspects. There are two main types of approaches to shallow parsing. One is base on rule-based methods; the other based on statistical methods. There is now a growing interest in applying machine-learning techniques to chunking, as they can avoid tedious manual work and are helpful in improving performance.

Much work has been done by researchers in this area. Li et al. used Maximum Entropy (ME) model to conduct Chinese chunk parsing [1], Zhang and Zhou used the inner structure and lexical information of base phrases to disambiguate border and phrase type [2]. Zhou et al. introduced the Chinese chunk parsing scheme and separated constituent recognition from full syntactic parsing, by using words boundary and constituent group information [3]. Zhao and Huang systematically defined Chinese base noun phrase from the linguistic point of view and presented a model for recognizing Chinese base noun phrases [4]. The model integrated Chinese base noun phrase structure templates and context features. These studies achieved promising results. However, comparing Chinese shallow parsing performance is difficult because those papers use different chunk definition and different data sets.

In this paper, we explore the practical issues in Chinese shallow parsing and present results on Chinese shallow parsing using Conditional Random Fields (CRFs).

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 167–176, 2005.

CRFs [5] are models proposed recently that have the ability to combine rich domain knowledge, with finite-state decoding, sophisticated statistical methods, and discriminative, supervised training. In their most general form, they are arbitrary undirected graphical models trained to maximize the conditional probability of the desired outputs given the corresponding inputs. This method has been successfully applied in many NLP fields, such as POS tagging [5], noun phrase segmentation [6], Chinese word segmentation [7], named entity extraction [8] and Information Extraction [9][10].

In what follows, first, we briefly describe the general framework of Chinese shallow parsing and explore the practical issue in Chinese shallow parsing. Then, we describe CRFs, including how to conduct parameter estimation. Finally, we present experimental results and draw conclusions with possible future directions.

2 Chinese Shallow Parsing

Shallow parsing is the process of identifying syntactical phrases in natural language sentences. Several types of chunks – phrases that are derived from parse trees of Chinese sentences by flattening down the structure of the parse trees - provide an intermediate step to natural language understanding.

The pioneer work of Ramashaw and Marcus [11] has been proved to be an important inspiration source for shallow parsing. They formulate the task of NP-chunking as a tagging task where a large number of machine learning techniques are available to solve the problem. Therefore shallow parsing can be regarded as of as a sequence segmentation problem in which each word is a token in a sequence to be assigned a label. Without loss of generality, let $\mathbf{X}$ be a set of word sequences and $\mathbf{Y}$ be a set of syntactic labels. The training set is then a sequence of pairs of the form $(X1, Y1), (X2, Y2) \ldots (Xn, Yn)$, where $Xi \in \mathbf{X}$, $Yi \in \mathbf{Y}$. On the basis of such a training set, a shallow parser could be trained, and then it can make predictions on future, unlabelled examples.

2.1 Chinese Chunk Definition

Chunks were first introduced by Abney [12], who used them for syntactic parsing. According to his definition, a chunk is the nonrecursive core of an intra-clausal constituent, extending from the beginning of constituent to its head, but not including post-head dependents.

Like the definition of English chunk given by Abney, we define Chinese chunk as a single semantic and non-recursive core of an intra-clausal constituent, with the restriction that no chunks are included in another chunk.

To be able to represent the whole hierarchical phrase structure, 10 types of Chinese chunks are defined. The phrase categories are listed below, each followed by a simple explanation and an example.

Table 1. The Chinese chunk categories

No	Category	Explanation	Example
1	VP	verb phrase	[ADVP 先后/ad] [VP 颁布/vv 实行/vv 了/as]
2	DP	determiner phrase	[DP 这些/dt] [NP 经济/nn 活动/nn]
3	ADJP	adjective phrase	[ADJP 对内/jj 和/cc 对外/jj] [NP 政策/nn]
4	QP	quantifier phrase	[VP 增长/vv] [QP 一成/cd 至/cc 两成/cd]
5	FRAG	fragment phrase	[FRAG（/pu 完/vv ）/pu]
6	NP	noun phrase	[NP 德国/nr 政府/nn 发言人/nn] [NP 福格尔/nr]
7	PP	preposition phrase	[PP 在/p ２月/nt ２６日/nt] [VP 举行/vv]
8	LCP	phrase formed by “LC”	[VP 是/vc] [LCP 近年/nt 来/lc]
9	ADVP	adverbial phrase	[ADVP 已经/ad 或/cc 正在/ad] [VP 研究/vv]
10	CLP	classifier phrase	[QP 二十五/cd] [CLP 米/m] [NP 口径/nn]

To represent Chinese chunks clearly, we use 3 types of chunk border tags in this paper.

1. B-XP XP ∈ {VP, DP, ADJP, QP, FRAG, NP, PP, LCP, ADVP, CLP} denotes that the current word is the first word of chunk XP.
2. I-XP XP ∈ {VP, DP, ADJP, QP, FRAG, NP, PP, LCP, ADVP, CLP} denotes that the current word is inside of chunk XP.
3. O denotes that the current word is outside any chunk.

Using these chunk border tags, we can consider the Chinese shallow parsing as a tagging task.

2.2 Independency Assumption

HMMs learn a generative model over input sequence and labeled sequence pairs. While enjoying wide historical success, standard HMMs have difficulties in modeling multiple non-independent features of the observation sequence. They are generative, in the sense which that they represent a joint probability distribution P（X，Y）. Because this includes a distribution P（X） over the input features, it is difficult to use arbitrary, overlapping features while maintaining tractability.

2.3 Label Bias

Classical probabilistic automata [13], discriminative Markov models [14], maximum entropy taggers [15], and MEMMs, as well as non-probabilistic sequence tagging and segmentation models with independently trained next-state classifiers [16] are all potential victims of the label bias problem [5]. This is because the per-state normalization requirement of next-state classifiers – the probability transitions leaving any given state must sum to one. Each transition distribution defines the conditional probabilities of possible next states given the current state and next observation element. Therefore, the per-state normalization requirement means that observations are only able to affect which successor state is selected, and not the probability mass passed onto that state which results in a bias towards states with low entropy transition and, in the case of states with a single outgoing transition, causes the observation to be effectively ignored [17]. In Chinese parsing, this problem is extremely severe.

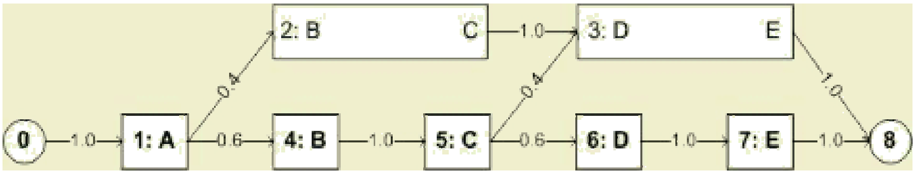

Fig. 1. Label bias problem

For example, Figure 1 represents a simple finite-state model designed to shallow parsing. The optimal path 0-1-4-5-6-7-8 is indicated by bold font. But the path 0-1-2-3-8 will have a higher probability and then be selected in decoding, because P (0, 1, 4, 5, 6, 7, 8|X) = 1.0*0.6*1.0*0.6*1.0*1.0 = 0.36, P(0, 1, 2, 3, 8|X)=1.0*0.4*1.0*1.0=0.4, P (0, 1, 4, 5, 3, 8|X) = 1.0*0.6*1.0*0.4*1.0 = 0.24, P (0, 1, 4, 5, 6, 7, 8|X) < P (1, 2, 3, 8|X) and P (0, 1, 4, 5, 3, 8|X) < P (1, 2, 3, 8|X). This is case that the states with a single outgoing transition effectively ignore their observations. More generally, states with low-entropy next state distributions will take little notice of observations.

3 Conditional Random Fields (CRFs)

CRFs are a recently introduced [5] from of conditional model that allow the strong independence assumptions of HMMs to be relaxed, as well as overcoming the label-bias problem exhibited by MEMMs [18]. This allows the specification of a single joint probability distribution over the entire label sequence given the observation sequence, rather than defining per-state distributions over the next states given the current state. The conditional nature of the distribution over label sequences allows CRFs to model real-world data in which the conditional probability of a label sequence can depend on non-independent, interacting features of the observation sequence. In addition to this, the exponential nature of the distribution chosen by Lafferty et al. enables features of different states to be traded off against each other, weighting some states in a sequence as being more important than others.

CRFs are defined as follows. Let $X = x_1 x_2 ... x_T$ denote some observed input data sequences, such as a sequence of words in training data. Let $Y = y_1 y_2 ... y_T$ be a set of finite state machine (FSM) states, each of which is associated with a label. By the Hammersley-Clifford theorem, CRFs define the conditional probability of a state sequence given an input sequence X

$$p(Y \mid X) = \frac{1}{Z_X} \exp(\sum_{i=1}^{T} \sum_{k} \lambda_k f_k (y_{i-1}, y_i, X, t)) \tag{1}$$

where Z_X is a normalization factor over all candidate paths. In other words, it is the sum of the "scores" of all possible state sequence.

$$Z_X = \sum_{y \in Y} \exp(\sum_{i=1}^{T} \sum_k \lambda_k f_k(y_{i-1}, y_i, X, t))$$

$f_k(y_{i-1}, y_i, X, t)$ is a feature function. The feature functions can measure any aspect of a state transition $y_{t-1} \rightarrow y_t$, and the observation sequence X, centered at the current time step t.

λ_k is a learned weight associated with feature f_k. Large positive values for λ_k indicate a preference for such an event, while large negative values make the event unlikely.

Given such a model as defined in Equ.1, the most probable labeling sequence for an input X is Y^* which maximizes a posterior probability.

$$Y^* = \arg\max_Y P_\lambda(Y \mid X) \tag{2}$$

It can be found with dynamic programming using the Viterbi algorithm.

In the case of the commonly used graph structure for modeling sequential data, the general form of Equ. 1 can be expanded to

$$p(Y \mid X) = \frac{1}{Z_X} \exp(\sum_{i=1}^{T} \sum_k \lambda_k f_k(y_{i-1}, y_i, X) + \sum_{i=1}^{T} \sum_k \mu_k g_k(y_i, X)) \tag{3}$$

Where each $f_k(y_{i-1}, y_i, X)$ is a feature of the entire observation sequence and the labels at position i and $i-1$ in the corresponding label sequence, each $g_k(y_i, X)$ is a feature of the label at position i and the observation sequence, and λ_k and μ_k are feature weights.

In this situation, the parameters λ_k and μ_k corresponding to these features are equivalent to the algorithm of HMMs transition and emission probabilities. Although it encompasses HMM-like models, the class of CRFs is much more expressive, because it allows arbitrary dependencies on the observation sequence [5].

3.1 Parameter Estimation

Given the parametric from of a CRF in Equ.3, fitting empirical distribution involves identifying the values of parameters λ_k and μ_k which can be estimated by maximum likelihood, i.e. maximizing the loglikelihood L_Λ – maximizing the conditional probability of a set of label sequences, each given their corresponding input sequence.

$$\begin{aligned} L_\Lambda &= \sum_i \log P_\Lambda(Y_i \mid X_i) \\ &= \sum_i (\sum_{t=1}^{T} \sum_k \lambda_k f_k(y, y, X, t) - \log Z_{x_i}) \end{aligned} \tag{4}$$

To maximize L_Λ, we have to maximize the difference between the correct path and those of all other candidates. CRFs are thus trained to discriminate the correct path from all other candidates, which reduces the inference of label bias.

Lafferty et al. introduced an iterative scaling algorithm for Equ. 4 L_Λ and reported that it was exceedingly slow. Several researchers have implemented gradient ascendant methods, but naïve implementations are also very slow, because the various λ and μ parameters interact with each other increasing one parameter may require compensating changes in others. McCallum 2003 employs the BFGS algorithm, which is an approximate second-order method that deals with these parameter interactions.

4 Experiments

We conducted experiments comparing CRFs to HMMs on Chinese shallow parsing. Also, we compared the performance of the model trained using CRFs from different training data size.

4.1 Experimental Setting

We use the Penn Wall Street Journal Chinese Treebank (LDC-2001T11) as experimental data. It consists of about 100K words, 325 Xinhua newswire articles on a variety of subjects. We consider each sentence to be a training instance, with single words as tokens. Sections 1-300 were used as training set, sections 301-325 was used as the test set. Table 2 summarizes the information on the dataset. Table 3 shows the detail information of training set and test set. In this experiment we only use the pos tag of the current word t_i and the current word w_i as features.

Table 2. The simple statistics on dataset

Information	Value
# articles	325
# sentences	4185
# words	100k
# chunk types	10
# chunks	62633

Table 3. The number of each chunk type in dataset

Type	Data set	Training set	Test set
VP	13619	13211	408
DP	1322	1275	47
ADJP	3132	3082	50
QP	4008	3735	273
FRAG	593	564	29
NP	26807	25782	1025
PP	3754	3618	136
LCP	1358	1305	53
ADVP	4893	4747	146
CLP	3147	2922	225

4.2 Evaluation Metrics

We measure the performance in terms of tagging accuracy, precision, recall and F-score, which are standard measures for the chunk recognition.

$$accuracy = \frac{\#\, of\ correct\ tagged\ words}{\#\, of\ words\ in\ test\ corpus}$$

$$recall = \frac{\#\, of\ correct\ words}{\#\, of\ words\ in\ test\ corpus}$$

$$precision = \frac{\#\, of\ correct\ words}{\#\, of\ words\ in\ system\ output}$$

$$F_{\beta=1} = \frac{2 \times recall \times precision}{recall + precision}$$

4.3 Experimental Results

We first report the overall results by comparing CRFs with HMMs. Table 4 shows the results on the dataset described before with the best results in bold. Compared with the result of the HMMs the result based on CRFs leads to an improved performance on most types of Chinese chunks, except ADVP, DP, LCP and PP chunks. The precision of ADJP is 1.42% lower than that of HMMs, but the FB1 is 0.18% higher than that of the HMMs and the recall of QP is 0.73% lower than that of HMMs, but the FB1 is higher than that of the HMMs, which shows that CRFs significantly outperforms HMMs.

Table 4. The results based on CRFs and based on HMMs

	CRFs accuracy: **91.90**			HMMs-bigram accuracy: 90.17		
	precision	recall	FB1	precision	recall	FB1
ADJP	87.04	94.00	**90.38**	88.46	92.00	90.20
ADVP	100.00	99.32	**99.66**	100.00	99.32	**99.66**
CLP	98.25	100.00	**99.12**	96.98	100.00	98.47
DP	97.92	100.00	**98.95**	97.92	100.00	**98.95**
FRAG	96.55	96.55	**96.55**	72.97	93.10	81.82
LCP	100.00	100.00	**100.00**	100.00	100.00	**100.00**
NP	83.79	80.53	**82.13**	78.35	77.66	78.00
PP	100.00	100.00	**100.00**	100.00	100.00	**100.00**
QP	84.46	91.58	**87.87**	82.08	92.31	86.90
VP	94.03	96.57	**95.28**	93.64	93.87	93.76
ALL	89.74	89.89	**89.82**	86.65	88.21	87.42

In the following experiment, we use 25 files as test set, but the training set range from 25 files to 300 files. Figure 2 shows the performance curve on the same test set in terms of the FB1, P and R measure with respect to the size of training data. We can see that precision, recall, and FB1 improve rapidly when the size of training set has not reached 75 folders. After that, the improvement slows down significantly. From this figure, we can see that more training data can help improve Chinese shallow parsing performance.

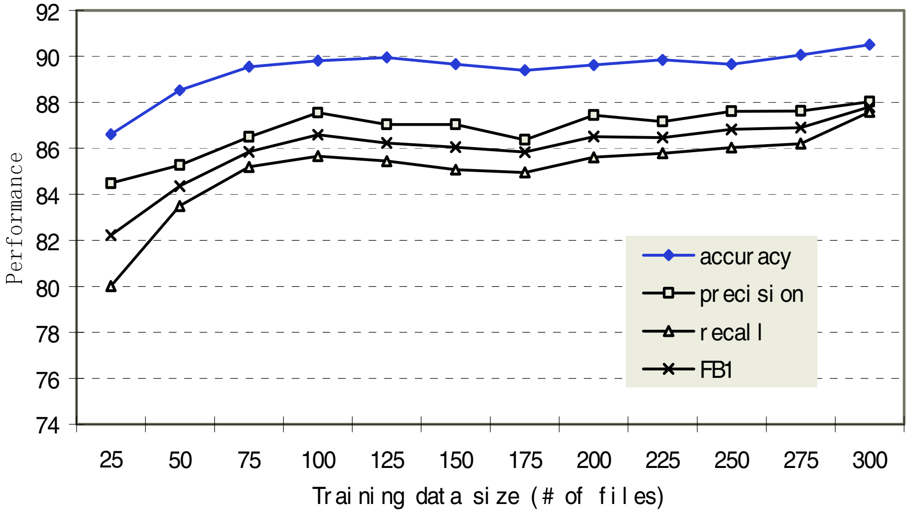

Fig. 2. The results based on CRFs vs. training set size

5 Discussion

From the experimental results we observed that the performance for Chinese shallow parsing do not look as good as those for English. One of the reasons might be that Chinese syntactic structure is more flexible and more ambiguous.

Looking through the errors in the results, we see that BAP, NP and QP internal structure is more flexible than another Chinese chunk type. The ambiguities of these syntactic structures lead to their poor experimental results. E.g. "[NP 支撑/nn 和/cc 推动/nn 作用/nn]" is likely to be selected compared to multiple tokens "[NP 支撑/nn 和/cc 推动/nn] [NP 作用/nn]" while the multiple tokens "[NP 美国/nr 网球/nn] [NP 公开赛/nn]" is likely to be selected compared to "[NP 美国/nr 网球/nn 公开赛/nn]". "[ADJP 老/jj] 、/ pu [ADJP 少/jj] 、/pu [ADJP 边/ jj] 、/pu [ADJP 穷/jj]" or "[QP ４ /cd] ：/pu [QP ３ /cd]" is likely to be selected compared to "[ADJP 老/jj 、/ pu 少/jj 、/pu 边/ jj 、/pu 穷/jj]" or "[QP ４ /cd ：/pu ３ /cd]".

Another question is that CRFs have many promising properties, but their main limitation is the slow convergence of the training algorithm relative HMMs, for which training on fully observed data is efficient.

6 Conclusion and Future Work

As far as we know the presented work is the first to apply CRFs to Chinese shallow parsing. In this paper, we present how conditional random fields can be applied to Chinese shallow parsing in which Chinese chunk boundary ambiguity exists. By virtue of CRFs, a number of correlated features can be incorporated and label bias can be minimized. We compare results between CRFs and HMMs in Upenn Chinese Treebank, and CRFs outperform the other approaches. Although we discuss Chinese shallow parsing, the proposed approach can be applicable to other language such as Thai.

From the experimental results we observed that more linguistic knowledge incorporated into the models may further improve the performance as well. Thus, our future work is to incorporate more contextual information into the models, including the boundary information of the phrases, semantic, collocation and co-occurrence information, aiming at further improvement of chunking in terms of the precision, recall and F score.

Another attractive aspect of CRFs is that one can implement efficient feature selection and feature induction algorithm for them. In the future we can start from features generating rules and evaluate the benefit of generated features automatically on data instead of specifying in advance which features of (X，Y) to use.

Acknowledgments. The authors would like to thank Cong Li and the anonymous reviewers for their helpful comments. This research was supported in part by National Natural Science Foundation (Beijing) (60083006), National 973 Project (Beijing) (G19980305011) and National Natural Science Foundation (Beijing) & Microsoft (60203019).

References

1. Sujiang Li, Qun Liu and Zhifeng Yang. Chunk Parsing with Maximym Entropy Principle. Chinese Journal of Computers. 2003. 26(12) 1734-1738.
2. Yuqi Zhang and Qiang Zhou. Automatic identification of Chinese base phrases. 2002. Journal of Chinese Information Processing, 16(16).
3. Qiang Zhou, Maosong Sun and Changning Huang. Chunking parsing scheme for Chinese sentences. 1999. Chinese J.Computers, 22(1):1158–1165.
4. Jun Zhao and Changning Huang. A transformation-based model for Chinese basenp recognition. 1998. Journal of Chinese Information Processing, 13(2):1–7.
5. John Lafferty, Andrew McCallum and Fernando Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 2001. Proc. 18th International Conf. on Machine Learning.
6. Fei Sha and Fernando Pereira. Shallow Parsing with Conditional Random Fields. 2003. In Proceedings of Human Language Technology-NAACL, Edmonton, Canada.
7. Andrew McCallum and Fang-fang Feng. Chinese Word Segmentation with Conditional Random Fields and Integrated Domain Knowledge.2003.
8. Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, Feature Induction and Web-Enhanced Lexicons. 2003. In Proceedings of Seventh Conference on Natural Language Learning (CoNLL).
9. David Pinto, Andrew McCallum, Xing Wei and W. Bruce Croft. 2003. Table extraction using conditional random fields. In Proc. of SIGIR, pages 235-242.
10. Fuchun Peng and Andrew McCallum. Accurate information extraction from research papers. 2004. In Proc. of HLT/NAACL.
11. L. A. Ramashaw and M. P. Marcus. Text chunking using transformation-based learning . 1995. Proceedings of the Third ACL Workshop on Very Large Corpora.

12. S. Abney. Parsing by chunks. 1991. In Robert C.Berwick, Steven P. Abney, and Carol Tenny, editors, Principle-Based Parsing: Computation and Psycholinguistics, Kluwer Academic Publishers, Boston, pages 257–278.
13. A. Paz. Introduction to probabilistic automata. Academic Press. 1971.
14. L. Bottou. Une approche theorique del'apprentissage connexionniste: Applications a la reconnaissance de la parole. 1991. Doctoral dissertation, Universite de Paris XI.
15. A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. Proc. EMNLP. New Brunswick, New Jersey: Association for Computational Linguistics.
16. V. Punyakanok. 2001. The use of classifiers in sequential inference. NIPS 13.
17. Hanna Wallach. Efficient Training of Conditional Random Fields. 2002. Thesis. Master of Science School of Cognitive Science, Division of Informatics. University of Edinburgh.
18. Andrew McCallum, D. Freitag and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. 2000. Proc. ICML 2000(pp. 591-598). Stanford, California.
19. Andrew Kachites McCallum. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass. edu. 2002.
20. Tianshun Yao, et al, Natural Language Processing – A research of making computers understand human languages, Tsinghua University Press, 2002, (In Chinese).