汉语 base NP 识别：错误驱动的组合分类器方法

徐昉，宗成庆

（中科院自动化化研究所模式识别国家重点实验室，北京，100080）

摘　要： 本文采用一种新的错误驱动的组合分类器方法来实现汉语 base NP(base noun phrase)识别。本文首先对汉语和英语 base NP 识别技术现状进行了综述和分析，明确了汉语 base NP 的含义，提出了错误驱动的组合分类器方法，然后，在对比两种不同类型的分类器——基于转化的方法和条件随机场分类结果的基础上，再利用支持向量机学习其中的错误规律，对两种分类器产生的不同结果进行纠错，从而达到提高系统整体性能的效果。在宾州汉语树库转化得到的 base NP 语料集上进行汉语 base NP 识别交叉验证的实验，与使用基于转化的方法，条件随机场以及支持向量机的方法相比较，实验结果都有所提高，F 值达到了 89.72%，相对于文中其他方法，最大提高了 2.35%。

关键词： 错误驱动，汉语 base NP 识别，组合分类器。

Chinese Base NP Chunking by Error-driven Combination System

Fang XU, Chengqing ZONG

(National Laboratory of Pattern Recognition, Institute of Automation,Chinese Academic of Sciences, Beijing, 100080, China)

Abstract: This paper proposes an error-driven combination approach to chunking Chinese base noun phrase (Chinese base NP), which combines TBL (Transformation-based Learning) model and CRF (Conditional Random Field) model. First, we gave a general overview and analysis of Chinese and English base NP chunking and present the content of Chinese base NP chunking. In order to analyze the result from two classifiers respectively and improve the performance of the base NP chunkers, an error-driven SVM (Support Vector Machine) classifier was designed to learn the errors found by comparison between the former two classifiers and modify those errors. According to our final experiments, our method achieves a higher accuracy in the final results with F-measure of 89.72% and improvement of 2.35% at most。

Keywords: Error-driven, Chinese Base NP chunking, Combination System.

1　引言

浅层句法分析，也称作部分句法分析（partial parsing）或语块（chunking）分析。浅层句法分析主要是识别句子中某些结构相对简单的独立成分，例如非递归的名词短语、动词短语等。这些被识别出的结构通常被称作语块（chunk）。它使句法分析的任务在某种程度上得到简化，同时也利于句法分析技术在大规模真实文本处理系统中迅速得到应用。现阶段最普遍的语块分析是 base NP 识别(base noun phrase chunking)。

Abney 首先提出了浅层句法分析的概念和策略，并设计和实现了一个简单的语块识别系统，此后，浅层句法

作者简介：徐昉（1983- ），男，安徽桐城人，硕士研究生，研究方向为 base NP 识别，fxu@nlpr.ia.ac.cn

分析，特别是 base NP 的识别得到了普遍的关注，国内外出现了很多 base NP 识别的方法，许多有效的统计和机器学习方法被应用到英语语块识别中，并且取得了比较理想的识别效果。Marcus [6] 第一次引入了机器学习的方法，将基于转化的错误驱动学习方法应用于 base NP 识别，做出开创性的工作，启发了以后的 base NP 识别工作。CoNLL-2000 (Conference on Computational Natural Language Learning 2000) 推出了英语语块识别的共享任务，同时该会议提供了大规模的英语语块库，为基于统计的英语 base NP 识别技术的研究提供了统一的训练和测试集。在会议 workshop 中的系统利用许多机器学习方法，其中效果最好的是 Kudo 和 Matsumoto [3]应用的支持向量机(Support Vector Machine)的方法。此后许多新的统计学习的方法应用到 base NP 识别中，例如条件随机场(Conditional Random Fields, CRF)[5]、结构学习方法（Structural Learning Methods）[1] 等等。Ando 和 Zhang[1] 提出了一种新颖的半监督学习的英文 base NP 识别方法并且取得了目前最好的识别结果。

汉语 base NP 识别还处于发展阶段，许多学者应用和英语 base NP 识别类似的方法来处理汉语 base NP，进行了有益的探索和卓有成效的工作。赵军[8]给出了汉语 base NP 的严格形式化定义，阐明了它的语言学内涵，提出了基于转换的汉语 base NP 识别方法和模型。还有许多其它的方法用于汉语 base NP 的识别，例如，隐式马尔可夫模型(Hidden Markov Model, HMM)[4]，最大熵(Maximum Entropy, ME)[10]，基于记忆的学习方法(Memory-based Learnig)[9]，等等。总体来说，汉语 base NP 识别速度和精度方面上还可以进一步的提高。

本文主要提出一种基于错误驱动的组合分类器方法，利用基于转化的机器学习方法和条件随机场的方法来识别汉语名词基本短语，在此基础采用错误驱动的策略，利用支持向量机来学习以上两个分类器对比所发现的一些错误规律，训练得到后处理纠错分类器，提高初级分类器产生结果的正确率。本文的其余部分是如下安排的：第二节给出了汉语名词短语的任务描述；第三节简单介绍实验中采取的模型；第四节为实验采取的识别算法及实现；第五节是实验设计和分析；最后一节是结论。

2 汉语 base NP 问题描述

base NP 指的是简单的，非嵌套的名词短语，不含有其它的子项短语[2]。Base NP 主要特性有两点：短语的中心语为名词；短语中不含有其它的子项短语，并且 base NP 之间结构上是独立的。赵军[8]从限定性定语出发给出了汉语 base NP 的严格形式化定义。典型的汉语 base NP 例如：甲级联赛、产品结构、自然语言处理等等。

3 模型介绍

Lafferty *et al.* [5] 提出了条件随机场（Conditional Random Fields, CRF）的概念，随后便被广泛地应用在模式识别各个领域， CRF 还被用作名词实体的识别，生物基因序列信息的识别等许多自然语言处理领域。CRF 模型描述如下。

给定的输出标识序列 Y 和观测序列 X，为了描述(X, Y)序列对上的 CRF，定义特征函数 $f_j(y_{i-1}, y_i, x, i)$ 和权值向量 λ， y_{i-1}, y_i 为标识序列， x 为输入序列， i 为输入位置。则

$$p(y \mid x, \lambda) = \frac{1}{Z(x)} \exp(\lambda g F(y, x))$$

$$F(y, x) = \sum_{i=1}^{n} f(y_{i-1}, y_i, x, i)$$

其中， $Z(x)$ 为归一化系数。

由上式求得条件随机场的条件概率，对于输入序列 x，最佳序列 y 可以通过下式确定

$$\hat{y} = \arg\max_{y} P(y \mid x) = \arg\max_{y} \lambda g F(y, x)$$

Sha 和 Pereira[7] 中提出 base NP 识别特征函数建立的方法。y_i 连续的标识序列为 $y_{i-1}=c_{i-2}c_{i-1}$，$y_i=c_{i-1}c_i$，特征函数 $f_j(y_{i-1},y_i,x,i)=p(x,i)q(y_{i-1},y_i)$

$p(x,i)$ 用来预测在当前位置 i 的输入标识 x，$q(y_{i-1},y_i)$ 预测输出 base NP 标识对。通过训练语料可以获得大量的 base NP 序列特征，通过训练可以得到 base NP 序列上的 CRF 模型。CRF 训练方法有共轭梯度方法，最小记忆类牛顿方法和投票感知器方法等。

SVM 是一种很有效的分类方法，主要思想是最大边缘原则和核函数原则。Kudo 和 Matsumoto [3]所实现的 SVM base NP 分类器利用训练语料的上下文信息——词，词性和 base NP IOB 标注来构建 SVM 向量空间，选取二维核函数，并且采取交叉验证和组合系统的方法取得了很好的识别效果。本文中主要选择 SVM 做为后处理分类器，将初级分类器的结果加入 SVM 分类器，从而提高识别结果，具体分析见第四节。

4 汉语 base NP 识别算法及其实现

机器学习中，组合分类器的方法可以提高学习效果。我们针对比较不同类型分类器所发现的识别错误，应用错误驱动的联合分类器方法修正错误，从而提高系统的识别性能。Zhou *et al.* [11] 指出基于 HMM 的 base NP 识别系统产生的错误主要是由某些词语引起，将这些词语加入 HMM 分类器重新训练，新的分类器可以提高 base NP 的识别结果。我们发现，CRF 和 TBL 对比得到的约 70%错误结果是由连续的 base NP 造成，错误的原因主要是由于 base NP 边界识别错误造成，和[11]相类似的方法，我们考虑利用支持向量机（Support Vector Machine，SVM）分类器来学习其中的错误规律。

实验分类器主要分为两个部分，初级分类器和后处理分类器。初级分类器是利用 TBL 和 CRF 训练得到的分类器来进行汉语 base NP 的识别；对比初级分类器的结果，利用一部分作为训练语料，将 TBL 和 CRF 识别的不同结果作为新的特征加入后处理分类器，并加入上下文信息，利用基于错误驱动的支持向量机的方法训练得到纠错分类器，然后对另一部分初级训练得到的结果进行测试，从而得到最终结果。实验系统主要流程如下：

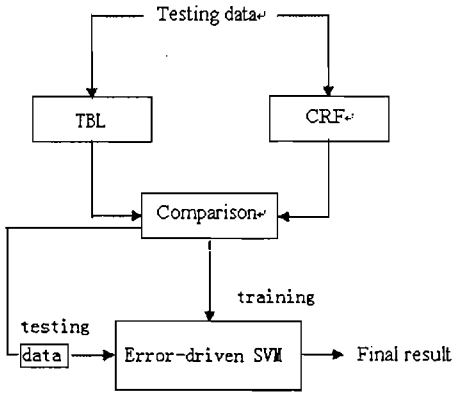

图1 实验主要流程

5 实验设计及分析

CoNLL 2000 提供了软件 chunklink[1] 将 Penn English Treebank II 转化成为名词短语识别所规定的 IOB 标注形式。本文采用实验数据来自宾州汉语树库 5.0，它比 Penn English Treebank II 含有更为复杂的词性标注集和树库标注。我们针对宾州汉语树库的新特性对 chunklink 做了一些修改，并加上后期一些手工标注处理，将

[1] http://ilk.kub.nl/~sabine/chunklink/

宾州汉语树库 5.0 转化为名词短语识别所需要的 IOB 标注形式。

实验中选取 CRF++[2]作为 CRF 分类器，fnTBL[3]做为 TBL 分类器，SVM 分类器采用工具 YacmCha[4]，其中 SVM 算法选择了 2 维多项式核函数，选定正参数 C=1；松弛变量 。

我们采取了三个标准来评价汉语 base NP 识别系统的性能，分别为精度（P）、召回率（R）和 F 量度（F），定义如下：

$$P=\frac{number\ of\ correct\ proposed\ baseNP}{number\ of\ proposed\ baseNP}*100\%$$

$$R=\frac{number\ of\ correct\ proposed\ baseNP}{number\ of\ corect\ baseNP}*100\%$$

$$F_{\beta}=\frac{(\beta^2+1)RF}{\beta^2R+F}\quad(\beta=1)$$

宾州汉语树库的总体大小约为 13MB，其中含有 505,144 汉语词，含有 base NP 个数为 14,1887。我们选取宾州汉语树库中 35 万汉语词作为初级分类器的训练语料，余下的 15 万汉语词作后处理分类器训练以及测试语料。为了保证实验结果公正合理，同时为了减小实验语料数据对结果的影响，在后处理分类器中我们采取交叉验证的方法，将 15 万汉语词分为 6 个相同大小的子集，用其中的 5 个子集做训练，余下的一个作为测试集，循环 6 次，使得每个子集都有机会最为训练集合测试集，实验得到结果如下：

	TBL			CRF			Error-driven SVM		
	P (%)	R (%)	F	P (%)	R (%)	F	P (%)	R (%)	F
Dataset1	87.10	88.11	87.61	89.43	87.86	88.64	90.29	88.88	89.58
Dataset2	87.47	88.53	87.99	90.01	88.10	89.04	90.49	88.76	89.62
Dataset3	86.30	87.35	86.82	89.12	88.10	88.61	90.17	89.24	89.75
Dataset4	87.71	87.29	87.50	87.87	87.37	87.62	88.43	88.21	88.32
Dataset5	86.77	87.81	87.29	88.94	88.21	88.57	91.23	90.91	91.07
Dataset6	86.44	87.58	87.00	89.59	88.35	88.96	90.55	89.37	89.96
average	86.97	87.78	87.37	89.16	88.00	88.57	90.19	89.23	89.72

表 1 三种模型在汉语 base NP 识别上的交叉验证实验结果

此外我们利用 SVM chunking 工具 YamCha, 与以上初级分类器相同的 35 万数据作为训练，在没有加入错误驱动信息的条件下，在 Dataset1 到 Dataset6 上测试得到的平均 F-measure 与 TBL，CRF 和 Error-driven SVM 三种方法相比较，图 2 给出四种模型在汉语 base NP 识别上的整体性能比较。

从表 1 和图 2 上可以看出，我们提出的 Error-Driven SVM 方法在汉语 base NP 识别上的性能最好，F 值为 89.72，主要是由于基于错误驱动的方法学习了初级分类器比较发现的一些错误规律，在 SVM 训练的过程中起到了纠错的作用；SVM 和 CRF 识别的结果比较接近，是由于 CRF 可以利用丰富的上下文信息，所以在识别连续的短 base NP 效果比较好，而 SVM 将模型向量空间映射到高纬空间，低纬空间的细颗粒特征贡献就相对减小，对于长度较长的 base NP 识别时 SVM 的效果很好。从图中可以看出基于统计的识别方法——SVM，CRF，SVM 和 Error-Driven SVM 要好于基于规则的识别方法，TBL 识别结果的 F 值较 Error-Driven SVM 模型低 2.35，的由于基于转化的方法不能利用训练数据中丰富的特征，通过训练得到的规则无法完全反映训练数据内在模型的规律，由于规则方法的局限性和歧义的存在，造成测试结果的不完备性。

[2] http://chasen.org/~taku/software/CRF++
[3] http://nlp.cs.jhu.edu/~rflorian/fntbl/
[4] http://chasen.org/~taku/software/YamCha

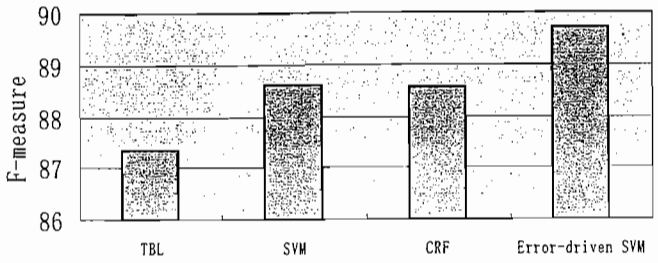

图 2：四种模型在汉语 base NP 识别上的总体性能表现

6 结论

本文采用一种新的基于错误驱动的组合分类器方法，融合了两种不同类型的分类器——基于转化的方法和条件随机场分类产生的共同正确结果，再利用支持向量机学习其中的错误规律，对两分类器产生的不同结果进行纠错，显著的提高了汉语 base NP 识别。本文系统的将几种主要的统计学习方法和基于转化的规则方法应用于大规模汉语 base NP 训练测试预料，并且比较了各种模型之间的实验效果，实验结果说明统计学习方法在大规模训练语料处理上比传统的基于规则的方法更加有效。如何利用新的机器学习方法和思想来解决 base NP 的识别问题，是我们下一步工作的重点。

参考文献：

[1] Rie Kubota Ando and Tong Zhang. A High-Performance Semi-Supervised Learning Method for Text Chunking. Proceedings of the 43rd Annual Meeting of ACL, 2005, 1-9.

[2] Church K. A stochastic parts program and noun phrase parser for unrestricted text. Proceedings of the Second Conference on Applied Natural Language Processing,1998

[3] Taku Kudo and Yuji Matsumoto. Chunking with support vector machine. Proceeding of the NAACL, 2001, 192-199.

[4] Heng Li, Jonathan J. Webster, Chunyu Kit, and Tianshun Yao. Transductive HMM based Chinese Text Chunking. IEEE NLP-KE 2003, Beijing, 2003, 257-262.

[5] J. Lafferty A. McCallum and F. Pereira. Conditional random Fields: probabilistic models for segmenting and labeling sequecne data. Proceedings of ICML,2001, 282-289

[6] Lance A. Ramshaw and Mitchell P. Marcus. Text Chunking using Transformation-Based Learning. Proceedings of ACL Workshop on Very Large Corpora, 1995, 82–94.

[7] Fei Sha and Fernando Pereira. Shallow Parsing with Conditional Random Fields. Proceedings of HLT-NAACL , 2003, 134-141.

[8] Zhao Jun and Huang Changling. A Quasi-Dependency Model for Structural Analysis of Chinese Base NPs. In: Proceedings of 36th ACL, 1998.

[9] Yuqi Zhang and Qiang Zhou. Chinese Base-Phrases Chunking. First SigHAN Workshop on Chinese Language Processing, COLING-02, 2002.

[10] 李素建, 刘群, 杨志峰. 基于最大熵模型的组块分析. 计算机学报, 2003,26(12):1722-1727

[11] GuoDong Zhou, Jian Su and TongGuan Tey. Hybrid Text Chunking. In: Proceedings of CoNLL-2000 and LLL-2000, 2000.