

# 基于最大熵模型的汉语基本块分析技术研究

李超<sup>1,2</sup> 孙健<sup>3</sup> 关毅<sup>1,2</sup> 徐兴军<sup>1,2</sup> 侯磊<sup>3</sup> 李生<sup>1</sup>

1. 哈尔滨工业大学计算机学院语言技术研究中心 150001

2. 哈工大-阿里巴巴联合实验室 150001

3. 阿里巴巴集团研究院 100026

E-mail: [beyondlee2008@yahoo.cn](mailto:beyondlee2008@yahoo.cn), [jian.sun@alibaba-inc.com](mailto:jian.sun@alibaba-inc.com), [guanyi@hit.edu.cn](mailto:guanyi@hit.edu.cn)

[xxjroom@163.com](mailto:xxjroom@163.com), [lei.hou@alibaba-inc.com](mailto:lei.hou@alibaba-inc.com), [lisheng@hit.edu.cn](mailto:lisheng@hit.edu.cn)

**摘要:** 本文论述了一个应用最大熵马尔科夫模型序列化标注块的边界、成分信息和应用最大熵模型分类识别块的关系信息的汉语基本块分析方法。为有效减少识别错误,重点探讨了候选标签筛选、难点关系识别等改进措施。集成上述方法的系统,边界、成分标记识别 F 值达到 93.196%,关系标记识别 F 值达到 92.103%,在中文信息学会句法分析评测(CIPS-ParsEval-2009)任务 2:汉语基本块分析中取得第一名。

**关键字:** 最大熵模型;最大熵马尔科夫模型;汉语基本块;块分析

## Chinese Chunking With Maximum Entropy Models

Li Chao<sup>1,2</sup> Sun Jian<sup>3</sup> Guan Yi<sup>1,2</sup> Xu Xingjun<sup>1,2</sup> Hou Lei<sup>3</sup> Li Sheng<sup>1</sup>

1. Research Center of Language Technologies, School of Computer Science and Technology,

Harbin Institute of Technology, Harbin 150001

2. Joint Laboratory between Harbin Institute of Technology and Alibaba.com 150001

3. Alibaba Group R&D 10026

E-mail: [beyondlee2008@yahoo.cn](mailto:beyondlee2008@yahoo.cn), [jian.sun@alibaba-inc.com](mailto:jian.sun@alibaba-inc.com), [guanyi@hit.edu.cn](mailto:guanyi@hit.edu.cn)

[xxjroom@163.com](mailto:xxjroom@163.com), [lei.hou@alibaba-inc.com](mailto:lei.hou@alibaba-inc.com), [lisheng@hit.edu.cn](mailto:lisheng@hit.edu.cn)

**abstract:** In this paper, we present a Chinese chunking method, in which boundary and composition identification is transformed into sequential labeling process by Maximum Entropy Markov Model, and relationship identification is transformed into classifying process by Maximum Entropy Model. Errors are significantly reduced by adding candidate tags selection and difficult relationship identification. The system integrated these methods achieved the F-measure of 93.196% for boundary and composition identification and the F-measure of 92.103% for relationship identification, and ranked the first in CIPS-ParsEval-2009 task2: Base Chunk.

**keywords:** Maximum Entropy Model; Maximum Entropy Markov Model; Chinese Base Chunk; Chunking

## 1 引言

组块是句子内部非递归的核心成分<sup>[1]</sup>。块分析是目前自然语言处理研究的重点问题,它既可以为作为从词法分析到完全句法分析的中间任务,又可以为机器翻译、信息检索、文本分类、问答等应用提供支持。文献[1]首先提出了块分析的思想,文献[2]提出了基于规则的块分析方法;2000年CoNLL推出了英文块分析块共享任务<sup>[3]</sup>,推动了英文块分析方法的研究,其中

占主导地位的是基于统计的方法，比较有代表性的有：隐马尔科夫模型<sup>[4]</sup>、最大熵模型<sup>[5]</sup>、支持向量机<sup>[6]</sup>、条件随机域<sup>[7]</sup>、winnow<sup>[8]</sup>、基于记忆的学习<sup>[9]</sup>、半指导学习<sup>[10]</sup>等。

在汉语组块体系建设方面，文献[11]对汉语基本短语进行了研究；文献[12]给出了基本名词短语的形式化定义；文献[13]建立了一个基于组块的句法分析器，提出11种组块的基本类型；文献[14]建立了一个完整的组块划分体系；文献[15]基于宾州大学中文树库给出了一套组块定义方式和组块类型。汉语块分析吸收了大量的英文块分析的方法，隐马尔科夫模型<sup>[6]</sup>、最大熵模型<sup>[17]</sup>、条件随机域<sup>[18]</sup>、基于规则驱动方法<sup>[19]</sup>、支持向量机<sup>[20]</sup>、基于大间隔方法<sup>[21]</sup>都被应用在汉语块分析之中。

文献[22]应用最大熵模型、最大熵马尔科夫模型及条件随机域进行基本块分析，是本文全部工作的基础。我们综合精度、效率、训练时间等各方面因素，选择了应用最大熵马尔科夫模型序列化标注块的边界、成分信息，应用最大熵模型分类识别块关系信息的汉语基本块分析方法。第2节将简介最大熵模型和最大熵马尔科夫模型。第3节阐述应用最大熵模型和最大熵马尔科夫模型进行块分析的细节及候选标签筛选、难点关系识别等改进措施。第4节将对本文工作做出总结和对下一步工作进行展望。

## 2 最大熵模型与最大熵马尔科夫模型

文献[23]在1957年基于香农信息熵理论建立了最大熵模型，文献[24]在1996年将最大熵模型引入自然语言处理中。其思想是：在信息不完全的条件下，一个存在不确定性的系统的最佳分布是使其信息熵最大的分布，很好的符合了“简单即最好”的哲学思想。假设  $h$  为上下文观测值， $t$  为标记，条件概率  $p(t|h)$  可以表示为：

$$P(t|h) = \frac{\exp(\sum_i \lambda_i f_i(t,h))}{Z(h)}$$

其中  $f_i$  为模型的特征， $Z(h) = \sum_t \exp(\sum_i \lambda_i f_i(t,h))$  为归一化因子。 $\lambda_i$  是特征  $f_i$  的权重，训练的过程就是求每个  $\lambda_i$  值的过程。

最大熵马尔科夫模型是最大熵模型的序列化形式<sup>[25]</sup>，以二元最大熵马尔科夫模型为例，将转移概率和发射概率合并到一个统一的条件概率函数  $P(t_i | t_{i-1}, h)$  中。通过加入能够表示先前标注信息的特征把  $P(t_i | t_{i-1}, h)$  转化成  $p(t|h)$  的形式求解。

最大熵马尔科夫模型蕴含了标记序列之间依赖关系的特点使其更加适合序列化标注的问题。当前词的边界、成分信息是与先前词的边界、成分信息密切相关，而短语关系识别更多的依赖于其内部词与词之间关系，因此我们采取分两步解决的块分析方法：1. 用最大熵马尔科夫模型序列化标注每个词的边界与成分信息，2. 用最大熵模型分类方法识别块的关系信息。

## 3 最大熵块分析及其改进措施

### 3.1 基于最大熵马尔科夫模型的基本块边界与成分分析

基本块分析的输入是词、词性信息序列，输出是基本块的边界、成分和关系信息。例：

输入：

执法/vN 部门/n 是/vC 反/v 腐败/a 斗争/vN 、/wD 搞好/v 廉政/vN 建设/vN 的/uJDE 重点/n 部门/n 之一/rN 。/wE

输出：

[np-ZX 执法/vN 部门/n ] [vp-SG 是/vC ] [np-ZX 反/v 腐败/a 斗争/vN ] 、/wD [vp-SG 搞好/v ] [np-ZX 廉政/vN 建设/vN ] 的/uJDE [np-ZX 重点/n 部门/n ] [np-SG 之一/rN ] 。/wE

在识别边界、成分时，我们利用最大熵马尔科夫模型采取 BIO 序列的方式对每个词进行标注，输入是词、词性信息，输出基本块的边界与成分信息。例：

输入：

执法/vN 部门/n 是/vC 反/v 腐败/a 斗争/vN 、/wD 搞好/v 廉政/vN 建设/vN 的/uJDE 重点/n 部门/n 之一/rN 。/wE

输出：

执法/vN/b-np 部门/n/i-np 是/vC/b-vp 反/v/b-np 腐败/a/i-np 斗争/vN/i-np 、/wD/o 搞好/v/b-vp 廉政/vN/b-np 建设/vN/i-np 的/uJDE/o 重点/n/b-np 部门/n/i-np 之一/rN/b-np 。/wE/o  
解码采取 viterbi 算法。通过实验对比我们得出三元最大熵马尔科夫模型比二元最大熵马尔科夫模型的效果更好，同时比较了各种特征窗口大小下系统的 F 值，发现以当前词为中心左右各三个词的特征窗口效果最佳。因此，在选择原子特征模板时，我们选择了以当前词为中心的左右各三个词的词、词性信息以及前一个词的标注信息。此外还应用了词与词、词性与词性、词与词性、标签与词性等复合特征模板。

### 3.2 基于最大熵模型的基本块关系分析

经过边界和成分识别后我们得到基本块的边界和成分信息，下一步是在结果上识别块的关系信息。采用最大熵模型对基本块的关系进行分类的方法，输入输出如下：

输入：

[np 执法/vN 部门/n ] [vp 是/vC ] [np 反/v 腐败/a 斗争/vN ] 、/wD [vp 搞好/v ] [np 廉政/vN 建设/vN ] 的/uJDE [np 重点/n 部门/n ] [np 之一/rN ] 。/wE

输出：

[np-ZX 执法/vN 部门/n ] [vp-SG 是/vC ] [np-ZX 反/v 腐败/a 斗争/vN ] 、/wD [vp-SG 搞好/v ] [np-ZX 廉政/vN 建设/vN ] 的/uJDE [np-ZX 重点/n 部门/n ] [np-SG 之一/rN ] 。/wE

经过最大熵马尔科夫模型的序列化标注边界、成分信息和最大熵模型分类标注关系信息后，得到含有边界、成分和关系信息的输出结果。在实现基本方法之后，我们又做了一些相应的改进，这里重点介绍候选标签筛选和难点关系识别改进。

### 3.3 候选标签筛选

在标注的过程中，我们发现性能受标注偏置影响<sup>[26]</sup>，此外，解码耗费时间较大。为此，我们首次提出了利用最大熵模型的候选标签筛选。筛选候选标签的方法是预留累进概率为某个阈值 T 内的排名靠前的若干个候选标签，其伪代码如下：

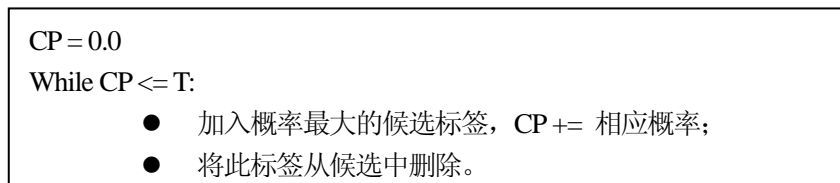


图1 候选标签筛选算法

这里经过实验得出 T=0.995 时效果最好, 这样就从所有的候选标签中筛选出最可能的部分, 通过排除不可能的标签在一定程度上减少了标注偏置的可能。

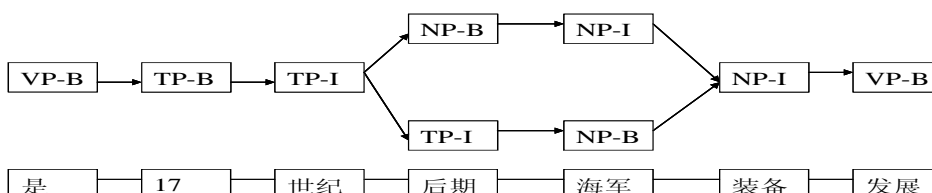


图2 候选标签筛选减少标注偏置问题的实例

图2中, 在我们的系统未加候选标签筛选时, 把“后期”和“海军”标注为 NP-B 和 NP-I, 但在经过候选标签筛选后标注为正确的 TP-I 和 NP-B。表1显示了边界、成分识别加入候选标签筛选后的最终结果和未加筛选相比较 F 值提升 0.1%。

表1 是否使用候选标签筛选结果对比

	准确率	召回率	F 值
未加筛选	92.960%	93.236%	93.098%
最终结果	93.041%	93.351%	93.196%

三元最大熵马尔科夫解码的时间复杂度为  $O(C^3 * L * S)$ , C 为平均候选标签的数量, L 为句子的长度, S 为给定观测序列及之前两个标签的情况下, 求当前标签的时间消耗。由于候选标签的减少(即 C 变小), 预测时间大幅减少: 2850 句清华评测语料的预测时间在 CPU 主频 1.6GH、内存 2G 的计算机上从八小时五十分变为 一分五十秒。

### 3.4 难点关系识别改进

假定块的边界、成分识别正确, 块的关系识别的难点在于右角中心结构 (ZX) 与链式关联结构 (LN) 的混淆、右角中心结构与并列关系结构 (LH) 的混淆。针对难点问题做了两点改进: 1. 针对 ZX 与 LN 混标的改进; 2. 针对 ZX 与 LH 混标的改进。

右角中心结构: 块中的所有词语直接依存到右角中心词, 形成一个右向中心依存结构。基本模式为:  $A_1 \dots A_n H$ , 依存关系为:  $A_1 -> H, \dots, A_n -> H$ 。H 为整个的句法语义中心词,  $A_1, \dots, A_n$  为修饰词<sup>[27]</sup>。

链式关联结构: 块中的各个词语依次依存到其直接右相邻的词语, 形成一个自左向右排列的多中心依存关系链。基本模式为:  $H_0 H_1 \dots H_n$ , 依存关系为:  $H_0 -> H_1, \dots, H_{n-1} -> H_n$ ,  $H_i$  成为不同层次的语义聚合中心,  $H_n$  为整个的句法语义中心词。

在文献[27]对基本块的成分和关系进行的统计中, 名词块占块总数的 50% 以上, ZX 和 LN 的名词块共占名词块的 94% 以上; 动词块的总数占块总数的 41% 以上, ZX 和 LN 的动词块共占动词块的 22% 以上。而且两种关系的结构又非常相像, 导致二者混标占关系识别错误的很大比

例(75.334%)<sup>1</sup>。因此有效地减少二者混标的错误，能够显著地降低关系识别错误率。

我们进行改进的思想是根据 ZX 和 LN 的定义构建特征模板建立最大熵模型区分器来区分二者。根据 ZX 应用块内每一个词和最后一个词的词、词性、词与词性信息作为其特征。类似地，我们定义块内每一个词和其后一个词的词、词性、词与词性信息作为 LN 的特征。我们把所有关系信息识别为“ZX”或“LN”的块都输入区分器中识别其关系结构。通过这种方法我们减少了 37.081% 的 ZX 和 LN 之间的错误识别，表 2 显示了关系识别在加入 ZX 和 LN 区分器后的最终结果与未加区分比较，F 值提升 0.42%。

表 2 是否使用 ZX 与 LN 区分器结果对比

	准确率	召回率	F 值
未加区分	91.531%	91.836%	91.683%
最终结果	91.950%	92.257%	92.103%

ZX 和 LH 之间的识别错误虽然没有 ZX 和 LN 之间的识别错误那样明显，但是仍占关系识别错误的 14.859%。因此，正确的区分二者同样有意义。与区分 ZX 和 LN 类似，我们建立一个最大熵模型区分器用于区分 ZX 和 LH。选取用块内每一个词和最后一个词的词、词性信息作为右角中心结构的特征，选取块内第一个词和其后每个词的词、词性信息作为并列结构的特征，把所有关系信息识别为“ZX”或“LH”的块都输入区分器中经过分类识别其关系结构。经过这种方法我们减少了 35.000% 的 ZX 和 LH 之间的错误识别，表 2 显示了关系识别在加入 ZX 和 LH 区分器后的最终结果与未加区分比较 F 值提升 0.08%。

表 3 是否使用 ZX 与 LH 区分器结果对比

	准确率	召回率	F 值
未加区分	91.872%	92.178%	92.025%
最终结果	91.950%	92.257%	92.103%

## 4 总结与展望

本文提出一种基于最大熵马尔科夫模型与最大熵模型的二步汉语基本块分析方法，它有以下特点：1.把边界、成分识别和关系识别分为两个过程，边界、成分识别用最大熵马尔科夫模型进行序列化标注，关系识别用最大熵模型进行分类；2.在关系识别中进行了候选标签的筛选，使 F 值提高，预测时间减少；3 通过建立区分器有效识别了难点关系结构。

条件随机域 (Conditional Random Fields) 是一种适合于序列化标注任务的无向图模型，已有学者应用条件随机域模型进行组块分析，我们将在下一步工作中应用此模型和最大熵模型进行对比分析。ZX 和 LN、ZX 和 LH 的区分器是二值分类器，而支持向量机(Support Vector Machine)在二值分类上效果明显<sup>[28]</sup>，我们将利用支持向量机模型改进区分器，并和现有的区分器比较。

<sup>1</sup> 此数据是在加入 ZX 与 LN 区分器前，系统输出与 Gold-standard 数据比较得出的结果。

## 参考文献

- [1] S. Abney. Parsing by Chunks. Kluwer Academic Publishers, Dordrecht. 1991:257-278
- [2] S. Abney. Part-of-Speech Tagging and Partial Parsing. Kluwer Academic Publishers. 1996:1-9
- [3] T. K. Sang, S. Buchholz. Introduction to the Conll-2000 Shared Task:Chunking. Proceeding of CoNLL-2000, bon, Portugal. 2000:127-132
- [4] A. Molina, F. Pla. Shallow Parsing using Specialized HMMs. Journal of Machine Learning Research. 2002, 2:595-613
- [5] R. Koeling. Chunking with Maximum Entropy Models. Proceedings of CoNLL-2000, Lisbon, Portugal. 2000:139-141
- [6] T. Kudoh, Y. Matsumoto. Use of Support Vector Learning for Chunk Identification, Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal. 2000:142-144
- [7] F. Sha, F. Pereira. Shallow Parsing with Conditional Random Fields. Proceedings of HLT-NAACL. 2003(5-6):134-141
- [8] T. Zhang, F. Damerau, D. Johnson. Text Chunking Based on a Generalization of Winnow. Journal of Machine Learning Research. 2002 (2):615-637
- [9] S. B. Park, B. T. Zhang. Text Chunking by Combining Hand-Crafted Rules and Memory-Based Learning. Proceedings of the 41st Annual Meeting of ACL, Sapporo, Japan. 2003:497-504
- [10] R. K. Ando, T. Zhang. A High-Performance Semi-Supervised Learning Method for Text Chunking. Proceedings of the 43th Annual Meeting of ACL, University of Michigan, USA. 2005:1-9
- [11] 周强. 汉语语料库的短语自动划分和标注研究. 北京大学博士论文. 1996: 1-9
- [12] 赵军, 黄昌宁. 汉语基本名词短语结构分析模型. 计算机学报. 1999, 22(2):141-146
- [13] M. Zhou. A Block-Based Robust Dependency Parser for Unrestricted Chinese Text. Proceedings of ACL-2000, Hong Kong, China. 2000:78-84
- [14] 张昱琪, 周强. 汉语基本短语的自动识别. 中文信息学报. 2002, 16(6):1-8
- [15] W. Chen, Y. Zhang, H. Isahara. An Empirical Study of Chinese Chunking. Proceedings of the 44th Annual Meeting of ACL, Sydney, Australia, 2006:97-104
- [16] 李珩, 杨峰, 朱靖波, 姚天顺. 基于增益的隐马尔科夫模型的文本组块分析. 计算机科学 2004, 31(2):152-514
- [17] Li SJ, Liu Q, Yeng ZF. Chunking parsing with maximum entropy principle. Chinese Journal of Computers. 2003, 26(12):1722-1727(in Chinese with English abstract)
- [18] Wenliang Chen, Yujie Zhang, Hitoshi Isahara. Chinese Chunking based on Conditional Random Fields. NLP2006, Yokohama, Japan. 2006:149-152
- [19] 周强. 基于规则的汉语基本块自动分析器. 第七届中文信息处理国际会议 2007:137-142
- [20] 李珩, 朱靖波, 姚天顺. 基于SVM 的中文组块分析. 中文信息学报. 2004, 18(2):1-7
- [21] 周俊生, 戴新宇, 陈家骏, 曲维光. 基于大间隔方法的汉语组块分析. 软件学报. 2009(4):870-877
- [22] 孙广路. 基于统计学习的中文组块分析技术研究. 哈尔滨工业大学博士学位论文. 2007
- [23] T. Jaynes. Information Theory and Statistical Mechanics. Physics Reviews. 1957(106): 620-630
- [24] A. Berger, S. A. DellaPietra, V. J. DellaPietra. A Maximum Entropy Approach to Natural Language

Processing. Computational Linguistics. 1996, 22(1):39-71

[25] A. McCallum, D. Freitag, F. Pereira. Maximum Entropy Markov Models for Information Extraction and Segmentation. Proceedings of ICML-2000, Stanford University, USA. 2000:591-598

[26] J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of ICML-2001, Massachusetts, USA. 2001:282-289

[27] 周强. 汉语基本块描述体系. 中文信息学报. 2007, 21(3) :21-27

[28] J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. 1998(2):121 – 167