

# 汉语基本块与功能块的自动分析

谷波<sup>1</sup>, 王瑞波<sup>2</sup>, 李济洪<sup>2</sup>

1. 山西大学 计算机与信息技术学院, 山西 太原 030006; 2. 山西大学 计算中心, 山西 太原 030006

Email: { gubo, wangruibo, lijh } @ sxu.edu.cn

**摘要:** 本文将汉语基本块和功能块的自动标注问题, 分别形式化为以词为标注单位的序列标注问题, 采用条件随机场模型进行了分析。模型选用不同窗口大小的词语层面的多个特征构造模型的特征模板, 使用正交表进行了最优特征模板的选择。在 2009 年的句法评测中, 本文的方法在汉语基本块的自动分析任务中 F 值达到 91.8%, 功能块的自动分析任务中达到 85.38%。

**关键词:** 汉语基本块; 汉语功能块; 条件随机场; 交叉验证; 正交表

## Automatic Labeling of Chinese Base Chunk and Chinese Functional Chunk

GU Bo<sup>1</sup>, WANG Ruibo<sup>2</sup>, LI Jihong<sup>2</sup>

1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006;

2. Computer Center of Shanxi University, Taiyuan, Shanxi 030006

Email: { gubo, lijh, wangruibo } @ sxu.edu.cn

**Abstract:** This paper formalizes Chinese base chunk labeling and Chinese Functional Chunk labeling as two models of sequence labeling at word-level, and employs conditional random fields to implement these two parsers. The candidate feature templates are constructed in terms of the different sizes of windows, and the best feature template is selected using the theory of orthogonal array. In the Chinese Parser Evaluation of 2009, our labeling model of base chunk can achieve the F-measure of 91.8%, and our model of functional chunk labeling achieves the F-measure of 85.38%.

**Keywords:** Chinese Functional Chunk; Chinese Base Chunk; Conditional Random Fields; Cross Validation; orthogonal Array

### 1. 引言

近年来, 中文信息处理技术得到了很大的发展, 在搜索引擎, 自动摘要, 以及问答系统等领域的得到了广泛应用。目前, 中文分词和词性标注已经达到了实用的价值, 但是汉语的句法自动分析性能仍不理想。本文认为主要的原因有两方面: (1) 中文短语分类体系缺乏统一的标准。(2) 中文缺乏大规模的标注的句法树库。为此, 清华大学周强教授提出了一套中文组块分析体系, 并建立了大规模的语料库来支撑该体系。这为推动汉语句法分析的发展奠定了重要的基础。本文针对汉语基本块和汉语功能块自动分析技术进行了研究。

传统的句法分析技术多是基于规则的方法。这种需要由语言学家给出或从语料库中自动获取规则, 然后在分析生句子的时候根据规则进行相应的标注。周强教授研制的汉语基本块自动分析

器<sup>[1]</sup>以及詹卫东教授研制的汉语完全句法分析器<sup>[2]</sup>均是采用了规则的方法。这种方法虽然可以有效地融合语言学知识，但却具有几点明显地不足：(1)规则的获取比较难，特别是由语言学家自己构造规则比较费时费力；(2)规则的覆盖范围有限：对于那些规则无法覆盖的短语现象，该方法就没有办法处理。(3)当规则之间存在冲突或歧义时，规则的排歧方法就成了模型性能的关键所在。

随着大规模语料库的构建，统计机器学习理论被广泛应用于自然语言处理中。这种方法可以从大规模语料中自动训练统计模型，并使用该模型对未标注的汉语句子实例进行自动预测，从而实现汉语句法结构的自动分析。与规则的方法相比，这类方法相对简单。当语料资源充足的情况下，模型可以对未标注实例作出准确的预测。本文主要采用了机器学习的方法来自动标注汉语基本块和汉语功能块。

本文主要总结了在 CIPS-ParsEval-2009 评测中，我们单位所使用的汉语基本块分析模型和汉语功能块分析模型。从官方的评测结果来看，该汉语基本块分析模型的 F 值可以达到了 91.98% (boundary+type) 及 89.85%(boundary+type+relation)；汉语功能块分析模型的 F 值可以达到 85.38%。这为进一步分析模型的不足提供了重要的参考信息。

## 2. 汉语基本块及汉语功能块简介

基于 Abney 的块理论，结合汉语的特点，周强定义了汉语基本块的描述体系<sup>[3]</sup>，即：基本块 = 基本拓扑结构 + 句法形式描述 + 语义内容描述，并给出了相应的基本块标记集合。该体系认为，汉语基本块的主要特点是块内部的各个词语按照一定的句法关系组合在以一词为中心的结构上，并可以通过这个中心词来体现整个基本块的外部功能。

周强<sup>[4]</sup>使用语块分析的方法提出了一套汉语功能块的描述体系，给出主语块(S)、谓语块(P)、宾语块(O)、兼语块(J)、状语块(D)、补语块(C)、独立语块(T)和语气块(Y)共八种功能块类型。该体系认为，汉语功能块是定义在句子层面的句法成分，具有穷尽性和线性两种性质，即：句子中的每个实义词都应无遗漏地进入某个功能块，所有功能块都处于同一层次，既不交叉也不存在包含关系。

## 3. 汉语基本块分析模型及功能块分析模型

### 3.1 汉语基本块分析模型

汉语基本块自动分析的目标，是给定一条汉语句子，自动地识别出句子中每个基本块的边界，并为每个基本块自动标注相应的句法标记和关系标记。本文将该任务分解成如下三个子任务：

- (1) 边界识别：本文将基本块边界的识别看作是词层面的序列标注问题，并使用 IOB2 策略<sup>[5]</sup>来描述每个词在基本块边界中的位置。
- (2) 句法标记标注：本文在边界识别的基础上，对每个基本块的句法标记进行了识别。识别时，仍然采用词语作为标注的基本单位，并使用标记集合{B-X, I-X, O}来同时描述每个词语的基本块位置及所具有的句法标记信息。其中，BIO 为词语在基本块中的位置，X 代表相应的句法标记。

- (3) 关系标记标注：类似于句法标记标注阶段的做法，本文将关系标记的标注看作为基于词的序列标注问题，并使用{B-X, I-X, O}来对每个词语的关系标记信息进行描述。其中 X 代表相应的关系标记。

在三个子任务中，本文除了使用了词特征、词性特征外，还使用了一些标点特征以及与标点的相对位置特征及它们之间的组合搭配特征，具体见第 4 部分。需要说明的是，对于句法标记标注和关系标记标注两个子任务，本文还将边界识别的结果作为特征加入到各自的模型中。

### 3.2 汉语功能块分析模型

汉语功能块自动分析的目标，是给定一条汉语句子，自动地识别出句子中每个功能块的边界，并标注出相应的功能块标记。

本文将功能块的标注看作是词层面的序列标注问题，使用 IOB2 策略，设置标记集合{B-X, I-X, O}，来实现功能块的边界和类型的同步识别。标记集合中，除 O 外，每个标记均有两部分组成。第一部分代表词语在功能块中的位置，功能块的开始位置使用 B 进行标记，功能块的内部位置使用 I 进行标记；第二部分为功能块的类类别标记，其中，X 遍历所有的功能块类型标记，例如：S(主语)，P(谓语)等。对于功能块外的词语，统一使用标记 O 进行标注。

功能块的自动标注模型使用了词层面的词特征和词语特征以及它们之间的组合和搭配特征。

### 3.3 条件随机场模型

本文使用条件随机场模型<sup>[6]</sup>来自动地预测汉语基本块和功能块的边界及类型。条件随机场模型是目前用于序列标注的最好的统计机器学习模型之一。它消除了 HMM 的强度独立性假设，并解决了 MEMM 的标注偏执问题，显著地提高了序列标注性能。

在训练阶段，条件随机场模型使用最大化后验估计(MAP)来估计模型的参数。其中，模型中每个特征的权重参数被假设服从均值为 0，方差为 C（默认为 1.0）的正态分布。

在测试阶段，条件随机场模型使用 Viterbi 算法来自动预测出每条序列的标注结果。需要说明的是，在基本块的句法标记标注子任务和关系标记标注子任务中，本文强制将预测出来的边界信息转化成基本块的边界识别子任务的结果。

## 4. 特征的选择及优化

### 4.1 基本块模型的特征选择

在基本块分析模型中，本文使用了如下六种词层面的特征：

- 1.词特征：当前标注的词语本身；
- 2.词性特征：当前标注词语的词性；
- 3.词距离上一个标点的位置：当前标注词语与前一个标点中间隔词语的个数；
- 4.词在句子中的位置：标识当前词语为句子中的第几个词；
- 5.词后面的第一个标点：当前标注词语后面的第一个标点，如果没有设置为 NUL；

考虑到在一条汉语句子中相邻词及词性之间存在这依赖关系, 本文还使用了词特征及词性特征的组合特征, 并对每种特征设置了若干个候选窗口。具体见表 1:

表 1: 汉语基本块分析模型的候选特征模板

特征类型	候选窗口			
词	[0,0]	[-1,1]	[-2,2]	[-3,3]
词性	[0,0]	[-1,1]	[-2,2]	[-3,3]
相邻词的二元组	-	[-1,1]	[-2,2]	[-3,3]
相邻词的三元组	-	[-1,1]	[-2,2]	[-3,3]
相邻词性的二元组	-	[-1,1]	[-2,2]	[-3,3]
相邻词性的三元组	-	[-1,1]	[-2,2]	[-3,3]
词/词性	-	[0,0]	[-1,1]	[-2,2]
词距离上一个标点的位置	[0,0]			
词在句子中的位置	[0,0]			
词后面的第一个标点	[0,0]			
基本块边界特征	[0,0] (这个特征仅在句法标记及关系标记的标注中使用)			

表 1 中给出了本文的基本块分析模型的候选特征模板。显而易见, 通过不同的窗口搭配, 共可以形成  $4^7$  个特征模板。

由于候选特征模板较多, 本文不可能逐一训练并找出最优的特征模板。因此, 本文使用了正交表的方法进行最优特征模板的选择, 该方法最初被用到了汉语框架语义角色的自动标注<sup>[7]</sup>研究中。本文采用正交表  $L_{32}(4^9 \times 2^4)$  构造出 32 个候选特征模板, 然后使用 32 个特征模板来训练条件随机场模型并在开发集上进行测试。最后根据 F 值从中选取出一个最好的模板作为最优特征模板。

为了避免在一份数据上进行实验, 对结果造成的随机性影响, 我们按照各种基本块大致相同的比例, 将训练数据均等的切分为 5 份, 然后对这 5 份依次选择一份作为开发集合, 将其余 4 份合并作为训练集合, 这样针对 32 个模板分别得到了 5 份交叉验证实验的平均 F 值。模板的好坏按照其所对应的平均 F 值进行评判。

## 4.2 功能块模型的特征选择

在功能块自动分析模型中, 本文仅用了如下两种特征:

1. 词特征: 当前标注的词语本身;
2. 词性特征: 当前标注的词语的词性;

为了精细地刻画每个待标注词语, 本文还使用了上述两个词语的组合特征信息, 并且为每种组合特征设置了相应的窗口。具体见表 2:

表 2: 功能块模型的特征模板

特征类型	窗口大小
词特征	[-2, 2]
词性特征	[-2, 2]

词的二元组合特征	[-1, 1]
词性的二元组合特征	[-1, 1]
词性的三元组合特征	[-2, 2]

本文所选用的功能块特征模板主要是参考文献[4]。

## 5 实验结果及分析

### 5.1 汉语基本块分析模型的实验结果

本文通过 5 份交叉验证实验，分别获得了基本块边界识别、句法标记标注和关系标记标注所用的特征模板，具体如表 3 所示。

表 3: 基本块分析模型所使用的特征模板

特征类型	边界识别模板 (C=2.0)	句法标记标注模板 (C=1.0)	关系标记标注模板 (C=1.0)
词	[-1, 1]	[-1, 1]	[-1, 1]
词性	[0, 0]	[0, 0]	[0, 0]
相邻词的二元组	[-3, 3]	[-3, 3]	[-3, 3]
相邻词的三元组	[-1, 1]	[-1, 1]	[-1, 1]
相邻词性的二元组	[-1, 1]	[-1, 1]	[-1, 1]
相邻词性的三元组	[-3, 3]	[-3, 3]	[-3, 3]
词/词性	[-1, 1]	[-1, 1]	[-1, 1]
词距离上一个标点的位置	[0, 0]	[0, 0]	[0, 0]
词在句子中的位置	[0, 0]	[0, 0]	[0, 0]
词后面的第一个标点	[0, 0]	[0, 0]	[0, 0]
基本块边界特征		[0, 0]	[0, 0]

从表 3 种可以看出，三个子任务所使用的特征模板基本相同。唯一的不同之处是句法标记和关系标记两个子任务所使用的特征模板中加入了基本块的边界特征。另外，在边界识别实验中，本文使用的 MAP 的方差参数 C 为 2.0，而其余两个实验使用的参数为 1.0。使用三个模板在整个训练集上进行训练，并在测试集上进行测试，得到了如表 4 所示的实验结果。

表 4: 基本块分析模型的评价结果

结果	准确率	召回率	F 值
边界	93.07%	92.88%	92.99%
边界+句法标记	92.07%	91.89%	91.99%
边界+关系标记	90.52%	90.34%	90.43%
边界+句法标记+关系 标记	89.94%	89.76%	89.85%

表 4 中分别给出了四组不同的实验结果，其中最后一行的实验结果为基本块整体的预测性

能。从表 4 中可知，基本块的自动分析的难点在于边界识别阶段。给定边界的基础上，句法标记的标注和关系标记的标注并没有带来性能的显著下降。因此，边界识别是基本块下一步研究的重点所在。

在提交评测结果之后，本文针对基本块的分析进行了深入的研究。本文曾尝试将 IOB2 策略改为 IOBES 描述，并在原有候选特征模板基础上加入了如下的特征：

词前标点特征：当前标注词前的标点。

在此基础上，本文仍然使用正交表进行特征模板选优，并进一步调节方差参数 C。最终，所得的三个子任务的最优特征模板如表 5 所示。

表 5：基本块分析模型所使用的特征模板

特征类型	边界识别模板 (C=5.0)	句法标记标注模板 (C=70.0)	关系标记标注模板 (C=85.0)
词	[-1, 1]	[0, 0]	[-3, 3]
词性	[0, 0]	[-1, 1]	[-3, 3]
相邻词的二元组	[-3, 3]	[-1, 1]	-
相邻词的三元组	[-1, 1]	[-3, 3]	[-1, 1]
相邻词性的二元组	[-1, 1]	[-2, 2]	[-2, 2]
相邻词性的三元组	[-3, 3]	[-2, 2]	[-3, 3]
词/词性	[-1, 1]	[-1, 1]	-
词距离上一个标点的位置	[0, 0]	[0, 0]	-
词在句子中的位置	[0, 0]	-	-
词后面的第一个标点	[0, 0]	-	[0, 0]
词前面的第一个标点	[0, 0]	-	-
基本块边界特征		[0, 0]	[0, 0]

利用表 5 中的特征模板，本文在给定的测试集上测试得到如表 6 所示的实验结果。

表 6：基本块分析模型的实验结果

结果	准确率	召回率	F 值
边界	93.54%	93.50%	93.52%
边界+句法标记	93.11%	93.08%	93.09%
边界+关系标记	92.20%	92.16%	92.18%
边界+句法标记+关系 标记	91.82%	91.78%	91.80%

对比表 4 和表 6 中的实验结果可知，新特征的加入、描述策略的改变以及 C 参数的重新调优使得基本块分析模型的整体 F 值有着近 2.0% 的提高。

## 5.2 汉语功能块分析模型的实验结果

本文利用 4.2 节给定的特征模板进行功能块标注实验，得到如表 7 所示的实验结果。

表 7: 汉语功能块分析模型的评价结果

结果	准确率	召回率	F 值
边界	88.13%	87.13%	87.62%
边界+功能标记	85.86%	84.89%	85.38%

从表 7 中可以看出, 功能块的边界识别结果较低, 仅能达到 87.62%。这主要是由于功能块结构相对比较复杂, 尤其是宾语块和状语块。另外, 由于功能块语料中, 有些功能块的分布较少, 条件随机场模型的预测会产生偏执现象, 从而不利于出现次数少的功能块的预测。

## 6. 总结

本文主要研究了汉语基本块和功能块的自动分析任务。对于汉语基本块的自动分析, 本文将其分解成边界识别, 句法标记标注和关系标记标注三个子任务, 并分别建立模型进行预测。最后, 本文三种信息融合到一起形成完整的标注。从实验结果来看, 我们所提交的结果可以整体可以达到 89.85% 的 F 值。进一步的实验结果表明, 通过变换标注策略, 模型的性能可以进一步提升, 模型 F 值最终达到 91.8%。

对于汉语功能块的自动分析, 本文将其看作是词层面的序列标注任务, 并将功能块的边界和类型同时进行识别。由于时间仓储, 本文仅使用了他人总结出的模板进行实验。实验结果表明, 功能块自动分析的最终 F 值可以达到 85.38%。值得一提的是, 正交表的特征模板选择方法可以有效地借鉴到汉语功能块的自动标注任务中去。另外, 将基本块形式化成特征信息引入到功能块的自动分析任务中去也是未来的一个重要研究方向。

## 参考文献

- [1] 周强. 基于规则的汉语基本块自动分析器[C]:第七届中文信息处理国际会议论文集, p137-p142, 2007
- [2] 詹卫东. 面向中文信息处理的现代汉语短语结构规则研究[T]: 北京大学博士论文, 1999
- [3] 周强. 汉语基本块描述体系[J]:中文信息学报, 2007 第 3 期
- [4] 周强. 汉语功能块自动分析[J]:中文信息学报, p18-p24, 21(5):2007
- [5] Lance A. Ramshaw, Mitchell P. Marcus. Text chunking using transformation-based learning[C]. In Proceedings of the 3rd Workshop on Very Large Corpora, p88-p94, 1995
- [6] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data.[C] In Proceedings of the 18th International Conf. on Machine Learning, p282-p289. 2001
- [7] 李济洪,王瑞波,王蔚林. 汉语框架语义角色的自动标注研究进展[C]. 全国第十届计算语言学学术会议, 2009