

# 基于 CRFs 的汉语功能块分析

刘海霞, 黄德根, 周惠巍  
(大连理工大学 计算机科学与工程系, 辽宁 大连 116024)  
([huangdg@dlut.edu.cn](mailto:huangdg@dlut.edu.cn))

**摘要:** 提出一种基于条件随机域 (CRFs) 模型的特征模板优化策略识别汉语功能块, 对经过正确词语切分和词性标注处理的汉语句子进行功能块边界识别和功能信息标注, 从而得出功能块自动分析结果。在 2009 年 CIPS-ParsEval 的中文简体语料上进行了测试, 功能块识别的精确率、召回率和 F1-measure 值分别为 85.84%、85.07% 和 85.45%, 其中主语块、谓语块、宾语块和状语块四个典型功能块的 F1-measure 值分别达到了 85.16%、88.22%、81.75% 和 91.98%。实验结果表明, 在特征模板中加入更多上下文组合特征, 可以提高功能块的自动识别性能。

**关键词:** 汉语功能块; 条件随机域 (CRFs) 模型; 部分句法分析

## Chinese Functional Chunk Parsing Based On CRFs

LIU Hai-xia, HUANG De-gen, ZHOU Hui-wei

(Department of Computer Science and Engineering, Dalian University of Technology, Liaoning 116024, China)

**Abstract:** We focus on detecting the boundary of Chinese functional chunks and labeling the functional information in a sentence with correctly word segmenting and POS tagging. This paper proposes an approach that combines the feature template optimizing strategy with Conditional Random Field Model for automatic labeling Chinese functional chunks. Using test data from CIPS-ParsEval-2009, the precision, recall and F-1 measure of Chinese functional chunks reaches 85.84%, 85.07% and 85.45% respectively, of which the F-1 measure of subject, predicate, object and adverb functional chunk reaches 85.16%, 88.22%, 81.75% and 91.98% respectively. The experimental results indicate that the performance of automatic parsing of Chinese functional chunks is improved by extending the feature template with more contexts.

**Key words:** Chinese functional chunk; conditional random fields (CRFs) model; partial parsing

## 1 引言

汉语的功能块分析属于汉语句法分析范畴, 汉语的句法分析按其分析处理深度可依次分为词性标注处理、句法块分析、句法树分析等, 功能块分析就是句法块分析的一种。句法块分析作为一种部分分析技术, 其目标是对完整的句法树分析进行合理分解, 以达到提高分析效率和对真实文本处理的适应性和稳健性的目的。因此功能块分析作为一种较好的部分分析结果, 可以与完整的句法树分析结果有效配合, 形成可以适应不同应用需求的句法分析结果。

由于汉语功能块概念的提出和相应大规模语料库的开发<sup>[1]</sup>, 功能块分析成为了汉语句法分析研究的一个新的切入点。因其具有传统语块分析

的问题定义相对简单的优点, 同时又描述了汉语句子中的各个主要功能成分, 所以起到了为汉语句法分析和语义分析之间架起一道重要桥梁的作用<sup>[2]</sup>。从功能块的研究现状来看, 功能块自动分析问题的处理具有一定的难度。文献[3]以语块识别结果为基础自动构建德语和英语的部分句法树, 得到相应功能块自动识别的整体识别准确率分别为89.73%和90.04%, 召回率分别为61.45%和59.79%。文献[4]利用判定树模型进行各个功能块的边界识别研究, 得到了最高74.1%的F1-measure值。另外还有文献[2]使用了两种不同的功能块分析模型, 在词和词性的基础上利用CRF模型进行序列标注, 最终功能块整体识别的F1-measure值达到了78.63%。但是由于没有融入更丰富的词和词性上下文信息, 导致CRF模型对于一些复杂功

能块的识别性能较差。

为了进一步改善功能块自动分析器的性能，本文着眼于特征模板的制定和优化上，尽量融入更多的上下文词和词性信息，使用条件随机域模型（CRFs）对经过正确词语切分和词性标注处理的汉语句子进行功能块的自动识别。通过对功能块自身特点的深入分析和大量的特征选择和对比实验，表明在特征模板中加入涉及更多上下文的特征组合对汉语功能块自动识别性能的提高具有积极的影响。

## 2 功能块概述及建模分析

### 2.1 功能块概述

汉语功能块是定义在句子层面上的功能性成分，主要描述句子中反映不同事件内容的基本信息单元。他们一般占据了句子中的主语、谓语、宾语、状语、定语、中心语等功能位置，通过组合形成不同的事件句式，完成对真实世界中不同事件内容的再现描述，体现了汉语句子的基本骨架。汉语功能块分析的目的是正确标注出包括主语块、状语块、述语块、宾语块、补语块、兼语块、定语块、中心块、独立块、其他特殊块等功能块标记，以显示句子在小句层面上的基本结构及骨架。

汉语功能块定义具有如下性质<sup>[1]</sup>：1) 穷尽性：句子中的每个实义词都应无遗漏地进入某个功能块；2) 线性性：标注完成的功能块形成一个线性序列，即所有功能块处于同一层次，既不交叉也不存在包含关系。我们的目标是识别句子中的功能块信息，覆盖自顶向下进行事件句式拆分而形成的各个基本信息单元，形成进行进一步的事件骨架树分析的最小功能块描述序列。

### 2.2 建模分析

借鉴组块分析的研究经验，汉语功能块的自动识别也可以转化为一个序列标注问题<sup>[5]</sup>，而基于CRFs（条件随机域）<sup>[6]</sup>的机器学习模型可以任意添加有效的特征向量，是一种非常好的序列标注器<sup>[7]</sup>，所以我们选择使用CRFs模型来实现功能块分析的序列标注处理。通过为文本句子中的每

个词语标注一个合适的类别标记，实现功能块的自动分析。

为了标识块与块之间的边界和功能信息，我们采用IOB2的标注集合来标记功能块，标记集中的每个标记均由两部分构成，第一部分为词语在功能块中的位置，如功能块的起始位置用B表示，内部位置用I表示；第二部分为功能块的类型标记，主要包括S—主语块、D—状语块、P—述语块、O—宾语块、C—补语块、J—兼语块、A—定语块、H—中心块、T—独立块、X—其他特殊块等功能标记，然后在这两部分标记之间用“-”来分隔。对于不属于这几类功能块的单词和符号，我们统一使用O来标记。这样共有10种标记类型，加上O标记总共21种功能块标记。任意一个词被标记为21种功能块标记中的一种，标记为同一类别B和I的词，构成一个功能块，该功能块直到遇到下一个标记为B或者O的词为止。

以“能脱离其他运动形式独立存在”为例，用表1来展示功能块的标注举例，其中词语和词性两列是经过正确的词语切分和词性标注的，IOB2标记一列就是我们为每个词语标注的功能块标记。进而得到最后的标注结果为：[P 能/vM ] [P 脱离/v ] [O 其他/rN 运动/n 形式/n ] [P 独立/aD 存在/v ]。

表1 功能块的标注举例

词语	词性	IOB2 标记
能	vM	B-P
脱离	V	B-P
其他	rN	B-O
运动	N	I-O
形式	N	I-O
独立	aD	B-P
存在	V	I-P

## 3 基于CRFs的功能块分析

考虑到汉语功能块本身长度过长和组成结构复杂等特点<sup>[8]</sup>，我们将着眼点定在特征模板的选择上，除了包含丰富的词、词性特征、词与词性的组合特征，还尽量扩展其处理窗口的长度以及融入更多上下文词和词性的组合特征，然后用统

计的方法构建分析模型来达到更好的识别效果。为了进行分析对比实验，本文参考了文献[2]中的词和词性的特征组合编制了我们的特征模板一，再通过对功能块本身的深入分析和大量的特征选择实验，总结得出另一种特征组合方式的特征模板二，如下所示：

特征模板一：

- 1) 前后各三个词的词语和词性特征；
- 2) 相邻两个词的词性组合特征；
- 3) 次相邻两个词的词性组合特征；
- 4) 当前词的词性分别与前、后词的词语组合特征；

特征模板二：

- 1) 前后各两个词的词语和词性特征；
- 2) 相邻两个词的词性组合特征；
- 3) 次相邻两个词的词性组合特征；
- 4) 当前词的词性分别与前、后词的词语组合特征；
- 5) 相邻两个词的词性组合特征再分别与其正对应窗口为四的词语组合特征；
- 6) 后两个词的词性组合特征再分别与当前词、前词的词性组合特征。

同样以“能脱离其他运动形式独立存在”为例，用表2对应特征模板中所描述的词、词性及其上下文信息，其中位置一列既表示词与词之间的相对位置信息又表示特征模板所处理的窗口大小。

表2 特征模板的处理窗口举例

位置	词语	词性
POS:-3	能	vM
POS:-2	脱离	v
POS:-1	其他	rN
POS:0	运动	n
POS:1	形式	n
POS:2	独立	aD
POS:3	存在	v

## 4 实验结果与分析

### 4.1 实验数据及评价参数说明

本实验采用了清华大学提供的约48万词规模新闻学术类TCT（清华句法树库）语料库，充分利用TCT中提供的丰富句法标记信息，自动提取形成了相应的功能块标注语料库。该功能块标注语料库中的171个文件（共3.83M）用来作为训练数据，而提供给我们用于生成评测结果的测试数据大小为730K，这样训练集和测试集的大小比例约为5:1。

依据CIPS-ParsEval-2009（中文信息学会句法评测）的评测标准，对汉语功能块的标注结果进行评价的主要评价指标包括功能块分析的准确率（Precision,  $P$ ）、召回率（Recall,  $R$ ）和F-1测度（F-1 measure,  $F_{\beta-1}$ ）。以下给出了评价功能块识别性能指标的计算公式<sup>[7]</sup>：

(1) 功能块识别准确率（Precision）：

$$P = \frac{\text{正确功能块数}}{\text{召回功能块数}} \times 100\%$$

(2) 召回率（Recall）：

$$R = \frac{\text{正确功能块数}}{\text{功能块总数}} \times 100\%$$

(3) F-1测度（F-1 measure）：

$$F_{\beta-1} = \frac{2 \times P \times R}{P + R} \times 100\%$$

其中：

正确功能块数：某类正确分析的功能块总数

召回功能块数：某类自动分析的功能块总数

功能块总数：某类Gold-standard（标准的标注结果）功能块总数

### 4.2 实验结果及分析

本实验使用的工具为 Taku Kudo 开发的开源 CRF++-0.53 软件包，采用封闭的训练模式，即用事先制定好的两个特征模板，分别对训练数据进行训练，得到两个不同的功能块分析模型；然后用这两个模型分别对测试数据进行标注，得到功能块的自动标注结果。在实验过程中，我们发现：不同的特征模板会得到截然不同的功能块标注结果，其中包含单个词的功能块（总计 17694

个块)和包含多个词的功能块(总计 17551 个块)的识别效果也有很大的差别。

首先,我们单独考虑功能块分析的边界识别情况,也就是说,自动识别出的功能块的左右边界是否与 Gold-standard 标注结果完全一致。表 3 给出了用两个特征模板分别对包含单个词和多个词的功能块进行边界识别的结果比较。从中可以看出,使用特征模板二进行边界识别的准确率、召回率和 F1-measure 值分别达到了 88.13%、87.33%和 87.73%,其中的 F1-measure 值比使用特征模板一进行识别的结果提高了 0.57 个百分点。实验结果表明,融入更多上下文信息的特征模板二改善了功能块边界识别的性能,然而由于占功能块总数约一半的包含多个词的功能块的边界识别难度较大,所以对功能块的边界识别整体性能产生了不利的影响。

其次,我们综合考虑自动识别的功能块的左右边界和功能标记是否均与 Gold-standard 标注结

果完全一致。与只考虑边界识别情况不同的是,一个自动识别的功能块正确与否要看它的左右边界和功能标记是否全部正确。表 4 和表 5 分别列出了两个特征模板下对包含单个词和多个词的功能块的左右边界加功能标记进行识别的准确率、召回率和 F1-measure 值。可见,特征模板二对单词功能块进行识别得到的 F1-measure 值已经达到了 89.44%,相应地对多词功能块进行识别的 F1-measure 值为 81.42%;同时,对于单词功能块,使用特征模板二测试得出的 F1-measure 值比使用特征模板一测试得出的 F1-measure 值提高了 0.30 个百分点,而对于多词功能块相应的 F1-measure 值提高了 0.83 个百分点。从这些实验数据可以看出,我们已经能够识别出绝大多数包含单个词的功能块;并且相对于特征模板一,在特征模板二中增加的特征组合对于更复杂的多词功能块的识别效果有更加显著的提高。

表3 包含单个词和多个词功能块的边界识别结果比较

功能块类别	特征模板一			特征模板二		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
单词功能块	91.03	90.15	90.59	91.15	90.70	90.92
多词功能块	84.10	83.32	83.71	85.06	83.94	84.49
合计	87.58	86.75	87.16	88.13	87.33	87.73

表4 包含单个词的功能块的边界加功能标记识别结果比较

功能标记	特征模板一			特征模板二		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
D	94.86	96.68	95.76	94.87	96.88	95.87
P	88.42	90.88	89.64	88.55	91.76	90.13
S	89.08	87.90	88.49	88.60	87.94	88.27
J	90.37	84.72	87.46	88.57	86.11	87.32
T	84.48	78.40	81.33	87.07	80.80	83.82
O	86.16	73.80	79.51	88.56	74.00	80.62
H	77.48	65.46	70.97	77.46	65.16	70.78
C	82.35	37.84	51.85	100.00	37.84	54.90
A	67.74	20.69	31.70	72.86	25.12	37.36
合计	89.57	88.70	89.14	89.66	89.22	89.44

表5 包含多个词的功能块的边界识别加功能标记识别结果比较

功能标记	特征模板一			特征模板二		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
X	89.74	93.33	91.50	90.51	95.33	92.86
J	87.76	84.31	86.00	88.54	83.33	85.86
D	82.77	86.78	84.73	84.28	87.29	85.76
P	83.91	84.20	84.05	84.80	85.04	84.92
S	82.19	82.76	82.47	82.39	82.99	82.69
O	79.54	83.67	81.55	79.85	83.96	81.86
H(C)	64.45	44.24	52.46	75.00	49.09	59.34
C(H)	70.00	38.18	49.41	68.75	46.20	55.26
A	62.50	28.30	38.96	69.78	30.50	42.45
T	42.17	26.92	32.86	53.33	30.77	39.02
合计	80.96	80.22	80.59	81.96	80.88	81.42

最后，我们给出功能块自动分析的完整结果，将测试集中所有功能块统合在一起，列出全部功能标记在两种特征模板下识别得出的准确率、召回率和 F1-measure 值。下面表 6 是包括全部功能标记的完整分析结果，从中可见，利用特征模板二测试得出的准确率、召回率和 F1-measure 值分别比利用特征模板一测试得出的结

果高出 0.55、0.55 和 0.57 个百分点，同时，特征模板二下的状语块、述语块、主语块和宾语块的 F1-measure 值分别已经达到了 91.98%、88.22%、85.16% 和 81.75%。可以说我们在特征模板二中加入的特征组合起到了提高自动分析结果的作用，并且得到了 85% 以上的准确率、召回率和 F1-measure 值。

表6 全部功能标记的完整分析结果比较

功能标记	特征模板一			特征模板二		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
X	89.74	93.33	91.50	90.51	95.33	92.86
D	90.15	92.89	91.50	90.78	93.21	91.98
P	86.78	88.40	87.58	87.19	89.27	88.22
J	88.82	84.48	86.60	88.55	84.48	86.47
S	85.22	85.04	85.13	85.13	85.19	85.16
O	80.08	82.70	81.37	80.55	82.98	81.75
H	72.12	55.66	62.83	73.92	56.40	63.98
T	66.83	52.16	58.59	73.82	55.29	63.23
C	74.47	38.04	50.36	82.00	44.57	57.75
A	64.08	25.34	36.31	70.81	28.41	40.55
合计	85.29	84.48	84.88	85.84	85.07	85.45

## 5 结论

本文力图进一步改善汉语功能块自动分析的性能，提出了可以通过寻找有效的特征组合来优

化特征模板，在正确词语切分和词性标注的基础上，使用条件随机域模型进行汉语功能块的自动分析。从实验结果可以看出，该方法可以改善功能块的自动识别性能，并且在占功能块总数主

要份额的状语块（D 块）、述语块（P 块）、主语块（S 块）和宾语块（O 块）的识别上均取得了比较好的效果。但是由于宾语块及大多数包含多个词的功能块都具有长度长且组成结构相对复杂等特点，从而识别效果相对差一些。另外，那些占很小一部分份额的功能块比如其他特殊块（X 块）和兼语块（J 块）的识别效果虽然不错，但由于它们属于稀疏数据并不能对识别效果的提高起到决定性作用。因此如何进一步提高宾语块和其他多词功能块的识别准确率以及如何结合其他方法改善汉语功能块的整体识别性能，将成为我们后续研究工作的重点。

## 参考文献：

- [1] 周强, 任海波, 詹卫东. 构建大规模汉语语块库[A]. 黄昌宁, 张普主编自然语言理解与机器翻译[C]. 北京: 清华大学出版社, 2001.
- [2] 周强, 赵颖泽. 汉语功能块自动分析. 中文信息学报 [A], 2007, (21)5.
- [3] Sandra Kübler and Erhard W.Hinrichs. From chunks to function-argument structure: A similarity-based approach [A]. In: Proceedings of ACL/EACL 2001 [c]. Toulouse, France: 2001. 338-345.
- [4] Elliott Franco Dr bek, Qiang Zhou. Experiments in Learning Models for Functional Chunking of Chinese Text [A]. In: Proc. of IEEE International Workshop on Natural Language processing and Knowledge engineering[c]. Tucson, Arizona, 2001. 859-864.
- [5] Ramshaw L, Marcus M. Text chunking using transformation-based learning [A]. Proceedings of the Third ACL Workshop on Very Large Corpora[C]. Boston: 1995, 82-94.
- [6] J.Lafferty, A.McCallum, F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [A]. In: Proceedings of the 18th International Conference on Machine Learning [C]. San Francisco: Morgan Kaufmann, 2001, 282-289.
- [7] 黄德根, 于静. 分布式策略与CRFs相结合识别汉语组块. 中文信息学报. 2008.
- [8] 陈亿, 周强, 宇航. 分层次的汉语功能块描述库构建分析. 中文信息学报. (22)3, 24-31.