基于基本块的汉语功能块自动标注

李国臣[1] 王瑞波[1] 李济洪[2]

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: {ligc, wangruibo, lijh}@sxu.edu.cn

摘要： 本文研究了基于基本块信息使用条件随机场模型(CRF)对汉语功能块进行自动标注的问题。针对词和基本块的两种不同的标注策略，将汉语基本块信息分别形式化成相应的特征，通过大量的特征组合优化实验，进行特征选择和模型参数估计。实验结果表明，在CRF模型中，基本块相关特征信息的加入可以大幅度地提高功能块识别性能。开放测试表明，在以基本块为单位的标注策略下，功能块自动标注的F值达到89.12%，这是目前最好的汉语功能块自动标注结果。

关键词： 汉语基本块、汉语功能块、条件随机场模型

Automatically Labeling Chinese Functional Chunk Based On Base-Chunks

LI Guo-chen[1] WANG Rui-bo[1] LI Ji-hong[2]

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

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

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

Abstract: This paper presents a method of automatic labeling Chinese functional chunks, which uses the Conditional Random Field Model and is based on base chunk information. In view of two kinds of labeling strategies of word-based and base-chunk-based, Chinese base chunk information has been formalized into the corresponding features. Massive experiments have been performed on feature combination and parameter estimation. The experimental results indicate that, in the CRF model, the addition of base chunk correlation feature can obviously improve the performance of function chunk recognition. On the testing data set, under the labeling strategy based on the unit of base chunk, the F-measure of automatic labeling of functional chunks achieved 89.12%, which is the best result at present.

Keywords: Chinese base chunk; Chinese functional chunk; conditional random field model

0 引言

目前，国外针对英语的句法分析理论以及技术已日趋完善。其中，完全句法分析中HPSG[1]分析理论比较出名，它建立在CFG文法的基础上，扩充了特征结构和分析过程中的一套约束机制。穗志方等[2]使用信息熵来分析中完全句法分析中所使用的大量特征，并提出了一种有效的特征选取方法。段湘煜等[3]针对依存语法提出了一种基于动作建模的概率分析算法，并将这种方法应用到了10种语言上。实验表明，这种方法明显地改善了依存树分析的性能。在浅层句法分析方面，Abney提出了块理论，并开发了多层次的有限状态成分组块自动识别工具[4]。

汉语的完全句法分析方面，段湘煜等[5]将动作建模的分析方法应用进来，使汉语依存句法分析性能得到明显的提高。但是在问答系统、文本检索、语义分析等一些应用型的NLP任务中，由于目前自动完全句法分析结果精度不高，它的作用并不能很好地发挥出来。为了能够准确地自

动分析出汉语句子中的句法信息，周强[6]等结合汉语的特点和 Abney 的块理论，提出了汉语组块分析思想。他认为汉语句子中的每个词可以首先结合成基本块，然后，逐步向上结合，分别聚合形成多词块和功能块，并最终形成一个具有嵌套结构的汉语句法分析树。以此为理论基础，他构建了大规模的汉语语块库[7]。

汉语功能块作为汉语组块体系中的一个重要的组成部分，可以为句法分析和语义分析之间架起一道桥梁。为了有效地分析出一个汉语句子中的功能块信息，赵颖泽[8]等将功能块的识别看作一个序列标注问题，在词、词性的基础上，使用 CRF 模型进行序列标注，最终，功能块识别的标记 F 值达到了 78.63%。但是，这类仅使用了词和词性信息的功能块识别模型的性能已经达到极限。词、词性信息的不足和稀疏导致了模型对于一些复杂性的功能块识别性能较差。

进一步改善功能块自动识别的性能需要我们将一些结构性的信息加入到模型中。而汉语基本块刻画了汉语句子中实义词之间的拓扑结构和它们之间的聚合关系。这些信息可以区分出一句话中的主要成分和辅助性成分。基本块是汉语功能块的组成单位，我们认为它可以对功能块自动识别性能产生积极的影响。

本文在充分分析汉语基本块的描述体系的基础上，使用了不同的标注策略，将基本块不同侧面的信息融合到了功能块的识别模型中。实验结果表明：正确的汉语基本块信息可以明显地改善功能块识别的性能，而自动分析的基本块信息对汉语功能块性能的提升也有着积极的影响。

1 功能块识别的建模策略

虽然基本块和功能块分别刻画了汉语句子不同层面的信息，但是，它们之间却有着紧密的联系。基本块信息的加入，不但丰富了功能块识别模型所使用的信息，而且使建模策略也变得比较灵活。

1.1 功能块描述体系

赵颖泽[8]使用语块分析的方法给出了一套汉语功能块的描述体系。他认为汉语功能块是定义在句子层面的句法成分，它具有穷尽性和线性性两种性质。并定义了 8 种汉语功能块：主语语块(S)，谓语语块(P)，宾语语块(O)， 兼语语块(J)，状语语块(D)，补语语块(C)，独立语块(T)，语气块(Y)。在此基础上，赵颖泽对清华大学的 TCT 功能块语料库进行了统计，发现语料中 S、P、O、D 块所占的比例达到了 97%，因此，他在进行功能块标注任务时，仅对 S、P、O、D 四种语块进行识别。 根据这一统计数据，我们也将功能块的标注目标定位在 S、P、O、D 四种语块的识别上，而将其它块看作是功能块的块外词。

1.2 基本块的描述体系

周强[6]使用 Abney 的块理论，结合汉语意向性的特点，定义了一套汉语基本块的描述体系。他认为，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聚合到一个句法语义的中心词上，并通过这个中心词来体现整个句法块的外部功能。在构建基本块语料时，他根据块内词语数目的不同，将基本块分为单词块和多词块，并将多词块归纳到三种基本的拓扑结构中，即：左角中心结构（LCC）、右角中心结构（RCC）和链式关联结构（CHC）。在此基础上，他定义了一套基本块的完整的描述体系并给出了相应的标记集[6]。

1.3 汉语基本块与功能块的关系

汉语功能块和基本块是通过不同层面来描述汉语句子的结构信息的。基本块主要刻画了每个句子中关系比较紧密的实义词之间的聚合关系，而功能块是来描述句子层面上的功能性成分的。但是，它们都体现了一个句子中不同的词语组合所表现出来的一种结构性的信息。从理论上来讲上，一个汉语功能块是由几个不同的基本块组成的，它们之间是一种聚合的关系。

1.4 基于基本块的汉语功能块的标注策略

汉语功能块识别可以看作是一个序列识别问题。为了便于标注，我们需要很好地选取标注单位。在基于词的功能块识别任务中，由于只有词和词性信息作为标注使用的特征，所以，该功能块识别模型也只能选取词作为标注单位。由于汉语基本块也是功能块的组成单位，那么在基于基本块的功能块标注单位的选取上就变得比较灵活。本文为了将基本块有效地融合到汉语功能块识别模型中，决定采用如下两种标注策略：

（1）以词作为功能块的标注单位：在标注时，使用词、词性作为基础特征，然后将基本块信息转化到词的层面上进行融合。这种标注方法，没有改变标注单位的大小，只是将基本块信息通过特征的形式融合到模型中来。本文将比较这种策略[8]的实验结果，并分析基本块特征对功能块识别的影响。

（2）以基本块作为功能块的标注单位：由于功能块是定义在句子层面的，如果我们使用词语作为标注单位，对于结构较复杂的功能块，在有限的窗口内是很难容纳足够的信息来准确地判断每个词的功能块信息的。但是，如果我们在基本块层面进行标注，使用基本块中心词对句子信息进行有效地压缩，突显出句子的主干信息，这样更有利于复杂功能块的标注。

2 基本块信息的特征表示

汉语基本块主要包含拓扑结构、句法形式描述和语义内容描述三方面的内容[6]。但是，目前汉语基本块的语义内容描述无法很好地形式化出来。因此，我们仅使用了基本块的拓扑结构和句法形式描述两方面的信息，并将这两方面的内容通过以下几个侧面进行描述：（1）基本块的中心词及词性信息；（2）基本块的句法信息；（3）基本块的关系信息；（4）基本块的成分信息；（5）基本块的边界词及词性信息；（6）基本块的拓扑序列标记；

2.1 以词为标注单位的基本块特征描述

由于基本块和词是两种不同粒度的信息，我们无法直接将基本块信息融合到以词为标注单位的模型中，这就要求我们需要通过一定的方法将基本块信息转化到词层面上。IOB2策略[9]目前在组块分析中使用得比较广泛，它可以将块层面的信息有效地转化成词层面的信息。通过IOB2策略，我们便可以将基本块的各个侧面的信息转化成如下的形式：

（1）基本块的句法标记：{B-X，I-X，POS}。其中，B-X表示当前词是句法标记为X的基本块的开始词。I-X表示当前词是句法标记为X的基本块的第一个词后的其它词。如果当前词是基本块的块外词，那么我们将其词性作为它的句法标记信息。

（2）基本块的关系标记：{B-X，I-X，word}。其中，B-X表示当前词是关系标记为X的基本块的开始词。I-X表示当前词是关系标记为X的基本块的第一个词后的其他词。如果

当前词是基本块的块外词，那么我们将该词本身作为它的句法标记信息。

（3）基本块的成分信息：{B-X，I-X，O}。集合的定义类似于上述两种。其中，成分标记被定义为X。这里对待块外词，我们一律使用O标记。

（4）基本块中心词及词性信息：{Y,N,X}。我们分别使用三个标记来表示当前词是否是基本块的中心词。如果是则标为Y，如果不是则标记为N。如果当前词是基本块的块外词，那么我们将其标记为X。另外，我们可以将当前词的左右相邻的基本块的中心词信息加入到模型中。这部分信息可以由相邻块中心词的词形以及词性标记直接体现。

（5）基本块的边界词及词性信息：我们选取当前词的左相邻块的右边界词以及右相邻块的左边界词来表达出这类信息。这类信息的标记分别为该词以及词性本身。

（6）基本块的拓扑序列标记：{M, R, B, P, O, X, C, H, I, J, K, S, N}。通过这个标注集合，我们可以很容易地将每个基本块以及块外词的拓扑序列映射到词的层面上。

2.2 以基本块为单位的基本块特征描述

以基本块为标注单位进行功能块标注，通过提取基本块的中心词，将原始的句子压缩成一个只有基本块中心词构成的句子。这种做法可以将原始句子的主干信息有效地表达出来，更有利于一些复杂结构的功能块的标注。在以基本块为单位的标注策略中，基本块的中心词、词性以及基本块的句法信息、关系信息和成分信息可以直接使用它们的标记集合来表示。

由于我们没有找到好的方法将下层的词、词性信息通过一定的标记上升到基本块的序列中，因此不得不将每个词以及词性信息忽略。这也是这种策略的一个比较大的缺陷。

3 实验结果及分析

实验所用语料库来自于清华大学的所提供的 TCT 语料库。我们分别使用了其中的词、词性信息，人工标注的基本块信息，人工标注的功能块信息。并且，为了验证自动分析的基本块信息对功能块识别模型的影响，我们还使用了基本块自动分析器。我们将语料库中 185 个文件按 8：2 的比例进行了切分，其中训练集包含 148 个文件，测试集包含 37 个文件。

为了有效地评价功能块自动识别模型的性能，我们使用如下的块指标定义：

假设模型标注出的功能块总数为 Cp，其中正确的功能块（必须保证左右边界正确，并且功能块类型正确）数目为 Cc，在测试集中的 C 功能块的数目为 Co，那么相应的准确率、召回率和 F 值定义如下：准确率 P=Cc/Cp；召回率 R=Cc/Co；F 值 F=2*P*R/(P+R)。

我们使用 CRF 作为标注工具，使用 BIO2 策略对模型进行标注。

由于在赵颖泽[8]的大量实验中，已经证明了加入状态转移特征可以带来功能块识别性能的大幅度提高。因此，我们以下进行的所有实验中保留状态转移特征。

3.1 BASELINE 实验结果

我们仿照赵颖泽[8]的实验，以词为标注单位，使用他论文中最好的词和词性的特征组合，在我们的语料库上，进行了实验。为了方便对比，我们将这个实验结果作为我们的 BASELINE 结果，见表 1。

表 1 BASELINE 的实验结果（产生特征数：2812230）

	准确率	召回率	F 值
S	85.61%	80.78%	83.12%
P	85.16%	84.79%	84.97%
O	79.16%	82.74%	80.91%
D	84.41%	87.39%	85.87%
合计	83.83%	84.34%	84.08%

3.2 TOPLINE 实验结果

为了得到加入基本块信息后功能块识别性能的上界，我们使用了人工标注的基本块语料库。将正确的基本块信息作为特征加入到模型中，并不断地调节特征模板，使功能块的识别性能达到最优。

3.2.1 以词为单位的功能块识别性能

在词、词性信息的基础上，我们将基本块的句法特征、关系特征以及它们的二元、三元组合融入到模型中。得到如下的实验结果。

表 2：加入句法-BIO、结构-BIO 后的信息（产生特征数：3257064）

	准确率	召回率	F 值
S	88.02%	85.29%	86.63%
P	90.62%	91.55%	91.08%
O	82.56%	85.71%	84.11%
D	86.98%	89.96%	88.45%
合计	87.56%	88.81%	88.18%

从上表，我们可以发现：不管是在 baseline 模型中，还是加入了基本块信息，不同的功能块的自动识别性能是不同的，其中 P 块识别性能最好，这主要是因为 P 块中的平均词长度较少，而且结构简单；S、O 块长度较长，并且结构复杂，因此，它们的识别性能是最差的。

基本块句法特征和关系特征的加入可以明显的改善各类功能块的识别性能，使总体性能提高了 4%。

随后，我们将相邻块的边界词特征和中心词特征以及他们的搭配特征加入到模型中。得到了如表 3 所示结果：

表 3：加入相邻块的边界词、中心词信息（产生特征数：5470731）

	准确率	召回率	F 值
S	88.35%	85.53%	86.91%
P	91.02%	92.42%	91.71%
O	83.50%	87.11%	85.27%
D	87.52%	90.56%	89.02%
合计	88.09%	89.59%	88.83%

我们发现，相邻基本块中心词和边界词特征的加入可以对每类功能块的识别性能的提升起到积极的作用，但是，由于这些词特征比较稀疏，因此，它们的加入也会带来特征空间维数的剧增。这对模型的稳定程度没有益处。

3.2.2 以基本块为单位的功能块识别性能

在以基本块为单位的功能块标注中，我们无法将句中的每个词和词性融合到模型中，因此，

我们仅选用了中心词、中心词性、基本块的句法特征和关系特征四类特征以及它们的搭配来进行实验，得到了如下的实验结果：

表 4：以基本块为标注单位的实验结果（产生特征数：2705193）

	准确率	召回率	F 值
S	89.10%	86.04%	87.54%
P	91.34%	92.28%	91.80%
O	84.08%	87.01%	85.52%
D	87.90%	90.83%	89.34%
合计	88.56%	89.69%	89.12%

我们发现，使用基本块作为标注单位的标注策略仅使用少量的特征便可以达到令人满意的结果。

3.3 PIPELINE 实验结果

在开放测试中，是没有办法得到生句子的正确的基本块标注的。因此，为了得到一个使用的基本块标注模型，我们将训练集和测试集中的语料使用基本块自动分析器进行自动标注，并将自动标注的基本块信息作为特征。然后，通过实验来有效地分析和评价这种含有错误的特征对功能块自动识别的影响程度。

类似于 TOPLINE 实验，PIPELINE 实验的策略也可以分别使用词和基本块为两种标注单位。我们下面分别给出了它们的实验结果。

3.3.1 以词为标注单位的 PIPELINE 实验结果

我们在进行实验时，使用了 TOPLINE 系统中所归纳出来的最好的特征组合方式。然后，使用 CRF 模型，得到如表 5 所示的实验结果：

表 5：以词为标注单位的 PIPELINE 实验结果（产生特征数：5497182）

	准确率	召回率	F 值
S	86.31%	82.33%	84.27%
P	85.33%	85.95%	85.64%
O	79.72%	83.62%	81.63%
D	84.90%	88.74%	86.78%
合计	84.27%	85.57%	84.92%

从上表中我们可以看出，以词为单位的 PIPELINE 系统的性能介于 BASELINE 系统与 TOPLINE 系统之间的。我们认为，这主要是因为自动分析的基本块中含有错误的信息，这种信息致使了基本块的作用不能完全发挥。当然，这样的实验结果也是符合我们的预想的。

3.3.2 以基本块为标注单位的 PIPELINE 实验结果

同样地，我们使用以基本块为标注单位的 TOPLINE 实验中的最好的特征组合方式，进行了 PIPELINE 实验。实验的结果在表 6 中给出。

表 6：以基本块为标注单位的 PIPELINE 实验结果（产生特征数：2783700）

	准确率	召回率	F 值
S	85.53%	84.31%	84.92%
P	85.54%	88.64%	87.06%
O	81.44%	86.01%	83.66%

D	85.25%	89.76%	87.44%
合计	84.66%	87.61%	86.11%

从上表可以知道，以基本块为单位的 PIPELINE 系统的各个块的 F 值介于 BASELINE 系统和 TOPLINE 系统之间，这是符合我们的预想的。而且，就 PIPELINE 实验的两种标注策略来讲，以基本块为单位的标注性能要好于以词为单位的标注性能。这和 TOPLINE 实验中的结果也是一致的。

4 总结

本文中，我们将汉语基本块的信息形式化成不同侧面的特征，并有效地融合到了功能块的自动识别模型中。通过实验证明，正确的基本块信息可以很明显地改善功能块自动识别的性能。而自动标注的基本块信息由于精度上的问题，对功能块自动识别的性能的提升幅度介于加入正确的基本块信息和不加之间。

PIPELINE 实验的结果的提升不仅依赖于基本块自动预测的精度，也依赖于功能块模型中特征的使用策略。因此，如何排除自动预测的基本块信息的错误，更好地将基本块信息纳入到 PIPELINE 实验中是我们下一步的研究方向。

致谢

本论文所使用的 TCT 语料库以及基本块分析器，是由清华大学周强教授提供。并且，在本课题在选题及研究过程中，得到了清华大学周强教授的悉心指导。在此，向他表示诚挚的感谢。

参考文献

[1] Pollard, C.and I.A.Sag. 1994. Head-Driven Phrase Structure Grammar. Chicago: The University of Chicago Press.

[2]穗志方、赵军、俞士汶,统计句法分析建模中基于信息论的特征类型分析,《计算机学报》, 2001 年,第 24 卷第 2 期

[3]Xiangyu Duan, Jun Zhao, Probabilistic Parsing Action Models for Mul-ti-lingual Dependency Parsing. In Proceedings of CoNLL Shared Task Session of the Conference on Empirical Methods in Natural Language Processing and Computational Language Learning (EMNLP-CoNLL), June 28-30, 2007, Prague, Czech.

[4] Abney,S.Partial parsing via finite-state cascades.Natural Language ngineering,1996,2(4):337～344.

[5]段湘煜，赵军，基于动作建模的中文依存句法分析，《中文信息学报》，Vol. 21 No. 5, pp. 25-30，2007 年 9 月

[6]周强.汉语基本块描述体系 中文信息学报 2007 第 3 期

[7]周强，任海波，詹卫东 (2001). “构建大规模汉语语块库”，黄昌宁，张普主编《自然语言理解与机器翻译》，清华大学出版社，p102-107

[8]赵颖泽. 汉语功能块的自动分析 清华大学硕士论文 2006.12

[9] Lance A. Ramshaw and Mitchell P. Marcus. 1995.Text chunking using transformation-based learning.In Proceedings of the 3rd Workshop on Very Large Corpora, pages 88–94.