文章编号: 1003-0077(2007)05-0018-07

汉语功能块自动分析

周 强,赵颖泽

(清华大学计算机系 智能技术与系统国家重点实验室,北京 100084)

摘 要: 汉语功能块描述了句子的基本骨架,是联结句法结构和语义描述的重要桥梁。本文提出了两种不同功能块分析模型: 边界识别模型和序列标记模型,并使用不同的机器学习方法进行了计算模拟。通过两种模型分析结果的有机融合,充分利用了两者分析结果的互补性,对汉语句子的主谓宾状四个典型功能块的自动识别性能达到了 80 %以上。实验结果显示,基于局部词汇语境机器学习算法可以从不同侧面准确识别出大部分功能块,句子中复杂从句和多动词连用结构等是主要的识别难点。

关键词: 计算机应用;中文信息处理;汉语功能块;边界识别模型;序列标记模型;模型融合

中图分类号: TP391 **文献标识码**: A

Automatic Parsing of Chinese Functional Chunks

ZHOU Qiang, ZHAO Ying-ze

(State Key Laboratory of Intelligent System and Technology,
Dept. of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Chinese functional chunks are defined as a series of non-overlapping, non-nested skeleton segments of a sentence, representing the implicit grammatical relations between the sentence-level predicates and their arguments. In this paper, we proposed two statistical models for parsing four main functional chunks in a sentence. In the chunk boundary detection model, we focus on building the sub models based on SVM algorithm for detecting SP (subject-predicate) and PO (predicate-object) boundaries. In the sequence labeling model, we formulate the chunking task as a sequence labeling problem and base our model on CRF algorithm. By introducing some revision rules, we build a combined parsing model which integrates the advantages of both statistical models and have achieved the best F-Score of 82.93 %, 86.58 %, 78.46 % and 86.64 %for subject, predicate, object and adverb functional chunks respectively. Experimental results show that the complex clauses and serial verb structures are the main recognition difficulties.

Key words: computer application; Chinese information processing; functional chunk; boundary recognition model; sequence labeling model; model merging

1 引言

句法分析作为自然语言处理的基础技术,被广泛应用于信息抽取、信息过滤以及信息检索等研究领域。但由于自然语言所具有复杂性和不确定性,目前的完全句法分析器还无法达到令人满意的效果。因此目前句法分析技术的一个重要发展趋势就是避开现实存在的巨大障碍,由完全句法分析转向部分句法分析(Partial Parsing)的研究。

完全句法分析着眼于分析整个句子的句法信息,揭示句子中各个成分之间的完整句法关系。而部分句法分析则采用“分而治之”的策略,从句子的局部着手,降低分析的难度。基于语块(Chunk)的

收稿日期: 2007-04-15 **定稿日期**: 2007-06-25

基金项目: 国家自然科学基金资助项目(60573185;60520130299)

作者简介: 周强(1967—),男,博士,副研究员,主要研究方向为计算语言学、词汇语义学、机器学习;赵颖泽(1981—),男,硕士,主要研究方向为计算语言学。

句法分析属于部分句法分析,其目标是将具有语法关联的局部语境词语组合形成一个语块,从而为完全句法分析提供一个有效的中间结果。语块的概念最初源自认知心理学,后被引入计算语言学中,近年来得到广泛的关注和研究。从最初的基本名词短语识别[1],到一般语块识别[2]、小句识别[3]以及近期比较流行的浅层语义角色标注[4,5]研究,其研究内容和深度也在不断扩展。

汉语功能块概念的提出和相应的大规模语块库的开发[6]为汉语部分分析研究提供了一个新的切入点。它一方面禀承了传统语块分析的问题定义相对简单的优点,同时又因为其描述了汉语句子中的各个主要功能成分,使得它成为汉语句法结构和语义角色链接的重要桥梁。本文通过对功能块描述特点的深入分析,提出了两种不同分析模型: 边界识别模型和序列标记模型,并使用不同的机器学习方法进行了计算模拟。通过两种模型分析结果的有机融合,充分利用了两者分析结果的互补性,对汉语句子的主谓宾状四个典型功能块的自动识别取得了较好的实验结果。

2　问题分析

汉语功能块是定义在句子层面上的功能性成分,主要包括主语、谓语、宾语、兼语、状语和补语块等。它们体现了汉语句子的基本骨架。汉语功能块定义具有如下性质: 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。

这种自顶向下的功能块定义方式,可以比较方便地建立起句法层面的功能描述关系与语义层面的谓词论元关系的内在联系。但同英语常用的 Abney 语块定义[7]相比,汉语的功能块粒度更大,组成情况也更为复杂。因此自动识别的难度也更大。下面对本文主要使用的新闻类语块库的相关分布数据进行初步的定量分析。

实验语料库由 185 个新闻文件组成,总词数约 20 万。每个文件均由人工标注出功能块的类别和边界信息。由于新闻语料具有用语规范的特点,和口语相比,通常句式较长,结构也较复杂,具有一定的代表性,因此,我们认为选用新闻语料作为后续算法训练和测试的数据,具有很强的代表性,能够反映出问题的复杂性和一般语料处理过程中所具有的普遍性。

表 1　8 类功能块的基本分布数据

功能块标记	块总数	词语总数	平均长度
P-谓语	21 988	27 618	1.26
D-状语	19 795	46 919	2.37
O-宾语	14 289	61 401	4.30
S-主语	11 920	34 479	2.89
J-兼语	855	2 083	2.44
Y-语气	594	604	1.02
T-独立语	407	909	2.23
C-补语	244	444	1.82

表 2　S, O, D 块的长度分布

块长度	S 块数目	O 块数目	D 块数目
1	5 322	3 537	12 147
2	2 093	2 228	2 499
3	1 402	2 117	1 431
4	917	1 624	1 010
5	627	1 108	696
> 5	1 559	3 675	2 013
合计	11 920	14 289	19 796

表 1 显示了目前定义的 8 个功能块的基本分布数据。其中主谓宾状块分别占了块总数的 97 %和覆盖词语总数的 98 %,是我们研究的重点。从平均长度来看,O、S 和 D 块最为复杂,是我们处理的主要难点。表 2 进一步列出了 O、S 和 D 块的长度分布数据。从中可以看到,长度超过 3 个单词的 S 块占了 S 块总数的 26.03 %,其中长度超过 5 个单词的占 13.08 %。这些 S 块通常都由一些包含复杂定语从句的名词短语和完整的主语从句构成,它们会给自动分析过程带来很大的困难。在 O 块中,大约有 25.72 %的 O 块长度超过了 5 个单词,有 44.84 %的 O 块包含了 3 个以上的单词。这些 O 块往往包含一个结构相对完整的宾语子句,它们的构成更加复杂,识别起来也更为困难。虽然大多数的 D 块长度都小于 5 个单词,但由于许多 D 块可能由一些复杂的介词短语和方位结构构成,对它们进行准确识别的难度一点也不亚于复杂的 S 块和 O 块。

3 模型设计

3.1 边界识别模型

根据功能块的两条基本性质,可以把汉语功能块的分析问题看作是对一个句子进行若干次“切分”的过程,切分出的每一段就是一个完整的功能块,带有相应的功能块类型的标记。按照这个思路,可以将汉语功能块分析的问题转化为一个边界识别问题,形式化描述如下:

令 $S=\langle W, T\rangle$ 表示输入的句子,其中 $W=w_1 w_2 w_3 \cdots w_n$ 表示输入的词语序列,$T=t_1 t_2 t_3 \cdots t_n$ 为对应的词性标记序列。边界识别模型输出为一个功能块边界序列 $O=\langle B, P\rangle$。其中,$B=b_1 b_2 b_3 \cdots b_m$ 表示输出的功能块边界类型序列,每个 b_i 为二元组 $\langle C1, C2\rangle$,$C1$,$C2 \in \{S, P, O, J, D, C, T, Y\}$,$C1$ 表示该边界前的一个功能块的类型,$C2$ 表示该边界后的一个功能块的类型;$P=p_1 p_2 p_3 \cdots p_m$ 则表示与之对应的位置信息。为了保证合法的功能块输出,该边界序列应满足下列一致性条件:对于任意的 $1 \leq i \leq m-1$,$b_i=\langle C1, C2\rangle$,$b_{i+1}=\langle C3, C4\rangle$,有 $C2=C3$。下面给出一个具体处理实例,输入句子:

核电/n_1 是/v_2 一/m_3 种/q_4 安全/a_5、/、$_6$ 清洁/a_7、/、$_8$ 经济/a_9 的/u_{10} 能源/n_{11}。/。 (1)

其中总共包含 12 个词项(每个标点符号都作为一个词项),共有 11 个可能出现功能块边界的位置(使用数字下标标出)。如果功能块分析器能正确地对句子 S 进行分析,则在位置 1 和位置 2 处会分别出现一个 SP 和 PO 功能块边界。分析结果可以表示如下:

核电/$n_{\langle S,P\rangle}$ 是/$v_{\langle P,O\rangle}$ 一/m 种/q 安全/a、/、清洁/a、/、经济/a 的/u 能源/n。/。 (2)

据此,可以得到如下的功能块标注输出结果:[S 核电/n][P 是/v][O 一/m 种/q 安全/a、/、清洁/a、/、经济/a 的/u 能源/n]。/。 (3)

具体实现可考虑采用分治思想,将原有的复杂识别模型拆分为若干个简单二元分类问题,并逐一解决,最后加以综合。考虑到主谓宾块对句子意义描述的重要作用,本文主要对 SP 和 PO 边界识别子问题进行了深入研究。通过大量实验分析[8],从 Naïve Bayes 模型、决策树模型和 SVM[9] 中选择了最终性能最好的 SVM 模型作为后续处理的基础模型。

3.2 序列标记模型

功能块分析可以转化为序列标注问题,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。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,目前我们只考虑 P、D、O 和 S 这 4 类功能标记,这样可以尽量减小标记集的大小,从而减小算法的计算复杂度。这两部分标记字母之间使用“-”来分隔。对于不属于这几类功能块的单词,我们统一使用 O 来标记。这样,就得到如下含有 9 个元素的标记集:T={B-P, I-P, B-D, I-D, B-O, I-O, B-S, I-S, O}。

具体计算问题描述如下:设输入的序列为 $X=x_1 x_2 x_3 \cdots x_n$,其中 x_i 为一个词语,并带有相应的词性标记,如 x_i=核电/n,相应的输出序列为 $Y=y_1 y_2 y_3 \cdots y_n$,其中 $y_i \in T$。则对一个输入序列 X 进行标注的过程就是为其寻找一个最优的输出标记序列 Y 的过程。根据序列 Y 就可以方便地还原出相应的功能块标注结果来。例如:针对上面的输入句子(1),如果能得到以下的序列标记结果:“核电/n_{B-S} 是/v_{B-P} 一/m_{B-O} 种/q_{I-O} 安全/a_{I-O}、/、$_{I-O}$ 清洁/a_{I-O}、/、$_{I-O}$ 经济/a_{I-O} 的/u_{I-O} 能源/n_{I-O}。/$_{O}$”,我们就可以还原出上面同样的标注结果(3)。

考虑到 CRF(条件随机场)模型[10] 在各类序列标注问题,包括词类标注、专名识别、语义角色标注中都显示出了很好的处理效果,本文主要使用了 CRF 模型来实现功能块分析的序列标注处理。

4 实验结果

我们从实验语料库中随机抽取 18 个文件作为测试集,其余 167 个文件作为训练集。两者的词语总数分别为 27 888 和 179 516,分布比例约为 1∶9。在这个同样的数据集上,我们分别使用 SVM 模型和 CRF 算法进行了边界识别模型和序列标记模型的计算模拟。

4.1 边界识别模型实验结果

在 SVM 算法的实验中,我们采用 Cornell 大学的 Thorsten Joachims 开发的 SVM^{light} 软件包①。针

① http://svmlight.joachims.org/

对每个待识别边界位置,选择不同长度的观察窗口,形成了以下 4 个特征模板: T1: w-2w-1w1w2t-2t-1t1t2; T2: w-3w-2w-1w1w2w3t-3t-2t-1t1t2t3; T3: w-4w-3w-2w-1w1w2w3w4t-4t-3t-2t-1t1t2t3t4; T4: w-5w-4w-3w-2w-1w1w2w3w4w5t-5t-4t-3t-2t-1t1t2t3t4t5。

我们使用功能块边界的准确率、召回率和 $F_{\beta=1}$ 值来评价各个实验结果的优劣。定义如下: 设 Cp 表示算法所识别出的功能块边界总数,Cc 为识别出的正确的功能块边界总数,Co 为测试集中功能块边界的数目,则相应的准确率、召回率和 $F_{\beta=1}$ 值可以定义如下: 准确率 P = Cc/ Cp; 召回率 R = Cc/ Co; $F_{\beta=1}$ = 2 ×P ×R/ (P + R)。实验结果表明,在 T4 模板,两者的识别结果达到最佳: 对 SP 边界, P = 86.15 %, R = 68.89 %, F = 75.56 %; 对 PO 边界, P = 78.74 %, R = 86.12 %, F = 82.26 %。有关详细内容可参阅文献[11]。

SP 边界的识别结果和 PO 边界的识别结果存在着很大的不同,SP 边界的识别结果是准确率高于召回率,而 PO 边界的识别结果则是召回率高于准确率。二者之间的不同充分反映了两类问题的差异性,SP 边界可能往往比较模糊,算法在识别时,倾向于只识别那些比较明显的边界,而 PO 边界则由于宾语块较为复杂,容易出现宾语块被切碎的情况,而使得准确率低于召回率。

4.2　序列标记模型实验结果

我们使用 Taku Kudo 开发的开源 CRF++ 软件包 ver0.42①,进行大量特征选择实验[11],最终选择以下特征组合: 1) 左右各两个词的词语和词类信息; 2) 功能标记转移特征; 3) 相邻 2 个词语和词类的组合特征; 4) 相邻 3 个词语的词类组合特征。表 3 显示了 4 个主要功能块的性能分析结果。与上面的 SVM 结果评价方式不同,这里的评价对象是各个功能标记词语。定义如下: 设置算法标注出的功能块标记总数为 Cp,其中正确的标记数目为 Cc,在测试集中的 C 功能块的标记总数为 Co。而准确率、召回率和 $F_{\beta=1}$ 值的计算方法则与 SVM 模型相同。

表 3　SDPO 块的序列标记结果

功能标记	P(%)	R(%)	F(%)
S	85.59	75.71	80.35
D	88.42	86.44	87.42
P	87.54	86.61	87.07
O	85.28	93.02	88.98
合计	86.51	87.10	86.80

5　模型融合与汇总处理

边界识别模型和序列标记模型分别从两个不同角度进行了特定功能块的自动分析。可以预期,当不同的模型在识别结果上达成一致时,所得到的结果的准确率会比原有的单独模型更为可靠。并且通过两者实验结果的融合分析,可以使我们对功能块分析的主要难点有更全面的认识。

5.1　两个模型识别一致情况的分析

首先,我们将 CRF 的标记序列转换为 SP 和 PO 边界标注结果,并对 2 类模型识别一致的结果进行性能评价,得到了表 4 的实验数据。从中可以看出,2 类模型汇总结果的准确率有了较大提高,但其代价是整体召回率下降。因此,如果需要高精度的输出结果,使用 2 类模型的融合汇总结果是比较有效的。从平均结果来看,CRF 算法综合性能最优,如果对输出结果的精度稍低,也可以考虑直接采

表 4　两个模型的融合分析结果

	SP 边 界			PO 边 界		
模型	准确率	召回率	$F_{\beta=1}$	准确率	召回率	$F_{\beta=1}$
SVM	86.15 %	68.89 %	76.56 %	78.74 %	86.12 %	82.26 %
CRF	87.37 %	84.52 %	85.92 %	82.75 %	87.76 %	85.18 %
综合汇总	94.38 %	66.76 %	78.20 %	87.12 %	80.02 %	83.42 %

① http://chasen.org/～taku/software/CRF++/

用CRF算法输出的结果。其中我们最关心的两者一致结果中的分析错误,对SP和PO边界,分别占了6%和13%。它们是功能块自动分析的主要难点,对此的深入分析可以为后续的算法改进和完善提供重要的客观参考。

为了能够对各种错误类型的分布有一个初步的了解,我们从开放测试结果中随机抽取出了133个一致结果错误,通过人工分类分析,总结了以下常见错误类型:

(1) 复杂宾语块分析错误。这些宾语往往具有完整的句子结构,如果单独将宾语拆分出来分析,则算法的识别结果往往是正确的。但整体上的复杂宾语块却很难识别出来。如:

• [S 更为/dD$_{B\text{-}S}$ 严重/a$_{I\text{-}S}$ 的/u$_{I\text{-}S\ s}|_{p}^{1.25917}$[P 是/vC$_{B\text{-}P}$,/,$_{I\text{-}P\ p}|_{o}^{-4.51798}$[O左/f$_{B\text{-}S}$ 小腿/n$_{I\text{-}S}$-$\downarrow_{p}^{0.273332}$ 骨折/v$_{B\text{-}P}$ 了/u$_{I\text{-}P}$。/。$_{I\text{-}P}$ ①

• [P记得/v$_{B\text{-}S\ p}|_{o}^{0.461881}$[O 谭善和/nP$_{I\text{-}S}$ 同志/n$_{I\text{-}S}$ 作为/p$_{B\text{-}D}$ 工兵/n$_{I\text{-}D}$ 纵队/n$_{I\text{-}D}$ 特邀/vN$_{I\text{-}D}$ 代表/n$_{I\text{-}D}$ 参加/v$_{B\text{-}P}$ 了/u$_{I\text{-}P\ p}\downarrow_{o}^{1.1105}$ 成渝铁路/nS$_{B\text{-}O}$ 举行/v$_{I\text{-}O}$ 的/u$_{I\text{-}O}$ 开工/vN$_{I\text{-}O}$ 典礼/n$_{I\text{-}O}$,/,

在上述的第二个例句中,谓语"记得"后面跟随的宾语成分长达14个单词,此时CRF的标记结果倾向于以后面的宾语成分作为句子的主干,结果误将动词"记得"划入了S块中,而SVM的边界识别结果,则根据局部信息,准确地对P块和O块进行了切分,但是由于局部信息的局限性,同时也将O块误分为多个成分。

(2) 复杂定语从句识别错误,例如:

• [O任弼时/nP$_{B\text{-}S}$ 同志/n$_{I\text{-}S}$ 为/v$_{B\text{-}P\ p}\downarrow_{o}^{0.20386}$ 政治委员/n$_{B\text{-}O}$ 的/u$_{I\text{-}O}$ 红二方面军/nO$_{I\text{-}O}$。/。

• [S 仅仅/d$_{B\text{-}D}$ 读/v$_{B\text{-}P}$ 过/u$_{I\text{-}P\ p}\downarrow_{o}^{0.129868}$ 几/m$_{B\text{-}O}$ 天/qT$_{I\text{-}O}$ 三字经/nR$_{I\text{-}O}$ 百家姓/nR$_{I\text{-}O}$ 之类/rN$_{I\text{-}O}$ 的/u$_{I\text{-}O}$谭善和/nP$_{I\text{-}O}$ 同志/n$_{I\text{-}O}$

• [D 将/p$_{B\text{-}D}$ 原/d$_{I\text{-}D}$ 用于/v$_{B\text{-}P\ p}\downarrow_{o}^{0.30516}$ 本/rB$_{B\text{-}O}$省/n$_{I\text{-}O}$ 贫困/a$_{I\text{-}O}$ 地区/n$_{I\text{-}O}$ 的/u$_{I\text{-}O}$ 扶贫/vN$_{I\text{-}O}$ 基金/n$_{I\text{-}O}$

如果将"的"字前的成分单独抽取出来,我们可以看到算法的效果还是很不错的,但当我们需要语义信息,或是需要对文本内容的深入理解时,目前的机器学习算法也就无能为力了。这类错误也是所有错误中最难用机器来解决的。

还有一类复杂"的"字结构,将动词短语名词化,这类错误也很难以处理,例如:

• [S 持有期/n$_{B\text{-}S\ s}\downarrow_{p}^{0.433922}$ 超过/v$_{B\text{-}P\ p}\downarrow_{o}^{0.301208}$ 半/m$_{B\text{-}O}$ 年/qT$_{I\text{-}O}$、/、$_{I\text{-}O}$ 不/dN$_{I\text{-}O}$ 满/a$_{I\text{-}O}$ 一/m$_{I\text{-}O}$ 年/qT$_{I\text{-}O}$ 的/u$_{I\text{-}O}$,/,$_{I\text{-}O}$

• [D 其中/rN$_{B\text{-}S\ s}\downarrow_{p}^{0.18112}$[S 构成/v$_{B\text{-}P}$ 犯罪/v$_{B\text{-}O}$ 被/p$_{I\text{-}O}$ 依法/d$_{I\text{-}O}$ 追究/v$_{I\text{-}O}$ 刑事/b$_{I\text{-}O}$ 责任/n$_{I\text{-}O}$ 的/u$_{I\text{-}O\ s}|_{p}^{-1.8643}$[P 87/m$_{I\text{-}O}$ 人/n$_{I\text{-}O}$。/。$_{I\text{-}O}$

(3) 多个动词连用结构中的准确功能关系判定错误。例如:

• [P 好学/v$_{B\text{-}P\ p}\downarrow_{o}^{0.248725}$[P 上进/v$_{B\text{-}O}$,/,

• [P 谈心/v$_{B\text{-}P\ p}\downarrow_{o}^{0.194207}$[P 服务/v$_{B\text{-}O\ p}\downarrow_{o}^{0.0642943}$[P 送/v$_{I\text{-}O\ p}|_{o}^{0.395486}$[O 温暖/a$_{I\text{-}O}$,/,

• [D 也/d$_{B\text{-}D}$ 开始/v$_{B\text{-}P\ p}\downarrow_{o}^{0.804252}$[P 进入/v$_{B\text{-}O\ p}|_{o}^{0.465176}$[O 关键性/n$_{I\text{-}O}$ 时期/n$_{I\text{-}O}$。/。[S 补给/v$_{B\text{-}P\ s}|_{p}^{-1.80124}{}_{p}\downarrow_{o}^{1.00848}$[P重/a$_{B\text{-}O}$ 于/p$_{I\text{-}O\ p}|_{o}^{-0.837048}$[O 作战/v$_{I\text{-}O}$。/。

由于目前的识别模型没有使用更复杂的词汇语义信息,因此对上面的各个相邻动词之间的功能关系很难准确识别。

(4) DP结构被误识别为SP结构,也是一致错误中比较难以处理的问题。例如:

• [D 总体/n$_{I\text{-}S}$ 上/f$_{I\text{-}S\ s}\downarrow_{p}^{0.240366}$[P 是/vC$_{B\text{-}P\ p}|_{o}^{2.13968}$[O 稳定/a$_{B\text{-}O}$[Y 的/y$_{I\text{-}O}$,/,

在状语块中常常会出现一些名词短语,如时间名词短语,方位名词短语等,当这些短语后接动词时,通常形成主谓结构,但少数情况也可以确定为状语成分。这些情况的准确判定需要更多的语言学知识。

(5) 由于补语块、兼语块等类型的功能块未列入我们的研究范围,因此,有时补语块、兼语块等会被误识别为宾语块。例如:

• [D 可谓/d$_{B\text{-}D}$[P 难/aD$_{I\text{-}D}$ 觅/v$_{B\text{-}P\ p}\downarrow_{o}^{1.0989}$[C 至极/dD$_{B\text{-}O}$。/。$_{I\text{-}O}$

• [P 吸收/v$_{B\text{-}P}$ 了/u$_{I\text{-}P\ p}\downarrow_{o}^{0.984103}$[J 万县市/n$_{B\text{-}O}$约/d$_{I\text{-}O}$ 4/m$_{I\text{-}O}$ 万/m$_{I\text{-}O}$ 名/qN$_{I\text{-}O}$ 劳动力/n$_{I\text{-}O}$[P到/v$_{B\text{-}P\ p}|_{o}^{1.0842}$[O 广东/nS$_{B\text{-}O}$[P 工作/v$_{B\text{-}P}$。/。$_{I\text{-}P}$

(6) 由于对一些结构的不同理解造成的不一致情况。例如:

---

① 相关标记说明:单词右下方为CRF算法输出的序列标记,单词之间带有"|"的标记为SVM算法输出的边界标记:"|"两侧的字母表示功能块边界的类型,斜上方的数字为SVM算法进行分类时输出的计算结果。每个标记都可能有3种字体来表示:粗体,~~删除线~~和*斜体*。粗体表示程序输出的标记为正确的结果,~~删除线~~表示输出的结果与正确结果不同,斜体表示程序未能识别出该位置处的相应标记。

• [D 为/$p_{B\text{-}D}$ 藏民/$n_{I\text{-}D}$ [P 送/$v_{B\text{-}P}$ 医/$n_{B\text{-}O}$ 送/$v_{B\text{-}P}$ 药/$n_{B\text{-}O}$

• [P 帮助/vJ $Y_{B\text{-}P}$ [J 库区/n_O 人民/n_O [P 重建/$v_{B\text{-}P}$ 家园/$n_{B\text{-}O}$、/、发展/$v_{B\text{-}P}$ 生产/$v_{B\text{-}O}$。/。

在上述的 2 个例句中,人工标注语料中的谓语成分粒度较粗,自动识别算法在此基础上做进一步的切分是可以接受的。

表 5 的统计数据显示,1、2、3 类错误所占的比例达到 78 %,是目前算法的主要处理难点。其中第 1 类错误往往是由于功能块本身过长,算法在缺乏语义知识的情况下,经常将一个完整的功能块进行分断处理。第 2 类错误则由于缺乏对功能块中复杂定语的准确理解而误将它们拆散为多个功能块。第 3 类错误反映了汉语缺乏形态变化的多个动词连用情况对自动识别算法的影响。要解决这些问题,一方面需要引入更多的词汇语义知识,另一方面,对目前的功能块标注体系进行重新设计,针对一些复杂功能块描述进行分层次的内容拆分,形成粒度较小和功能特征更为明确的基本功能块,应该也是一条可行的改进思路。

表 5　一致错误的统计数据

错误类别	错误数	所占比例
1	44	33.08 %
2	42	31.58 %
3	18	13.53 %
4	5	3.76 %
5	13	9.77 %
6	11	8.27 %

另外,一些可以不同理解的不一致情况 6 也占了 8 %左右,再加上判断比较模糊的情况 4 和目前分析器没有处理的情况 5,表明对目前的自动分析器的性能评价结果存在一定程度的低估现象。有关内容应在后续处理中加以注意。

5.2　两个模型识别不一致情况的处理

通过上面的实例分析,我们发现两种不同的模型在处理同一个句子时,存在一定的互补性,即在某些条件下一种模型的结果表现较好,而在其他的情况下,另一种模型的表现更佳。如果能够发掘出这些规律,就可以在两个模型出现不一致结果时,采用合理的策略进行修正,进一步提高最终的识别结果。

从这个思路出发,我们手工总结了 6 条规则,使用 SVM 输出对 CRF 结果进行修正,使整体的 F-Measure 值提高了大约 0.37 个百分点。虽然改进不大,但其性能提高幅度已接近或超过纯粹通过增加特征而产生的统计模型性能提高效果[11],显示出统计和规则相结合方法的处理优势。目前,人工制定的规则数目还比较少,覆盖的范围也不是很广,如果能继续增加一些合理的排歧规则,或者利用词汇关联对信息进行自动选择评价,相信会进一步改善最终的输出结果。

6　完整功能块识别结果的分析评价

通过以上的模型融合和自动修正处理,可以将最终的 CRF 标注结果转换成完整功能块标注,并进行以下针对完整功能块的性能评价: 令 Cp 为识别出的功能块的数目,Cc 为识别正确(功能块类型和左右边界全部正确)的功能块数目,Co 为测试集中的功能块数目。

表 6　完整功能块的识别结果

标记	准确率	召回率	$F_{\beta=1}$
D	71.09 %	71.36 %	71.22 %
O	75.32 %	81.87 %	78.46 %
P	81.83 %	83.87 %	82.84 %
S	84.61 %	81.31 %	82.93 %
Sum	77.80 %	79.46 %	78.63 %

表 6 的实验结果显示,目前的分析器在功能块层面上基本可以达到 80 %左右的准确率和召回率。其中宾语块的识别性能较差是与它的本身复杂度密切相关的。但 D、P 块的识别结果偏低,却出乎我们的预期。通过对处理结果的深入分析,我们发现目前的人工标注语料对 D、P 边界处理存在着许多不一致现象,特别是多个副词状语与谓语动词连用情况: 可能把多个副词标注为一个 D 块,也可能把它们分别标注为多个 D 块;可能把谓语动词单独标注为一个 P 块,也可能把它们与左边的一个或多个副词合并标注为一个 P 块。这种情况对人的理解没有问题,但通过统计模型学习并识别处理后,就会产生许多自动识别的 D、P 块边界与人工标注结果不一致现象,从而导致相关性能指标的下降。而这些情况,是不应该判定为自动分析错误的。通过考虑以上语言表述模糊情况重新进行 D、P 块的性能评

价,我们得到了以下新的 F-measure 值: D—86.64 %,P—86.58 %。它们比较客观地反映了目前自动分析器的真实处理能力。

目前国内外在功能块层面上进行的相关研究不是很多。CoNLL-2001 曾提出一个英语子句识别任务[3],其目标是自动识别英语句子中的所有嵌套子句。考虑到这个问题的复杂性,他们把它拆分成三项子任务: 子句起点识别、终点识别和完整嵌套结构识别。其中最困难的第三项子任务基本上与我们定义的序列标记问题难度相当。当时最好系统的开放测试 F1 值为 78.63 %[3],后来,通过改进算法,将分析性能提高到了 80.44 %[12]。文献[13]使用基于存储学习(MBL)方法,在语块识别结果基础上自动构建德语和英语的部分句法树,完成相应功能块的自动识别,整体识别准确率分别达到了 89.73 %和 90.04 %,召回率分别为 61.45 %和 59.79 %。德爱礼在我们目前的功能块标注语料库上,采用判定树模型进行了各个功能块的边界自动识别研究,通过使用从北大语法信息词典中提取的各种特征进行参数训练,最好 F1 值达到了 75 %左右[14]。这些研究从不同侧面证明了功能块自动分析问题的处理难度。

7 结论

自顶向下的复杂功能块设计为汉语部分分析研究提出了新的挑战。本文通过对功能块描述特点的深入分析,提出了两种不同分析模型: 边界识别模型和序列标记模型,并使用不同的机器学习方法进行了计算模拟。实验结果显示: 基于局部词汇语境的 SVM 模型和 CRF 算法可以从不同侧面准确识别出大部分功能块,句子中复杂定语从句、主宾语从句和多动词连用结构等是主要的识别难点。如何应用更丰富的词汇语义知识,并重新设计分层次的功能块描述体系,进一步提高目前自动识别器的处理性能,将是我们下阶段的主要研究重点。

参考文献:

[1] Lance A. Ramshaw and Mitchell P. Marcus. Text Chunking Using Transformation-Based Learning [A]. In: Proceedings of the Third ACL Workshop on Very Large Corpora8 [C]. Cambridge MA, USA: 1995.

[2] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to CoNLL-200 Shared Task: Chunking [A]. In: Proceedings of CoNLL-2000 and LLL-2000 [C]. Lisbon, Portugal: 2000. 127-132.

[3] Erik F. Tjong Kim Sang and Herv D jean. Introduction to the CoNLL-2001 Shared Task: Clause Identification [A]. In: Proceedings of CoNLL-2001 [C]. Toulouse, France: 2001. 53-57.

[4] Xavier Carreras and Llus Marquez. Introduction to the CoNLL-2004 shared task: Semantic role labeling [A]. In: Proceedings of the Conference on Computational Natural Language Learning (CoNLL) [C]. Boston, MA: May, 2004.

[5] Xavier Carreras and Llu s M arquez. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling [A]. In: Proceedings of the CoNLL-2005 [C]. 2005.

[6] 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. 102-107.

[7] Steven Abney. Parsing By Chunks [A]. In: Robert Berwick, Steven Abney and Carol Tenny (eds.), Principle-Based Parsing [C]. Kluwer Academic Publishers, Dordrecht. 1991.

[8] Yingze Zhao, Qiang Zhou A SVM-based Model for Chinese Functional Chunk Parsing [A]. In: Proc. of the Fifth SIGHAN Workshop on Chinese Language Processing[C]. Sydney: 2006. 94-101.

[9] Vladimir N. Vapnik. The Nature of Statistical Learning Theory [M]. Springer, 1995.

[10] John Lafferty, Fernando Pereira, and Andrew McCallum. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [A]. In: International Conference on Machine Learning (ICML '01) [C]. 2001. 282-289.

[11] 赵颖泽. 汉语功能块的自动分析 [D]. 北京: 清华大学,2006.

[12] Xavier Carreras1, Lluis Marquez, et. al. Learning and Inference for Clause Identification [A]. In: Proc. of ECML '02 [C]. 2002.

[13] 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.

[14] 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.