基于条件随机域和语义类的中文组块分析方法

孙广路[1,2], 郎 非[3], 薛一波[1]

(1. 清华大学 信息技术研究院, 100084 北京, guanglu. sun@ gmail. com;
2. 哈尔滨理工大学 计算机科学与技术学院, 150080 哈尔滨; 3. 哈尔滨理工大学 外国语学院, 150080 哈尔滨)

摘 要: 为了解决中文组块分析精度不高和未利用词的语义信息的问题,提出了一种基于条件随机域模型和语义类的中文组块分析方法. 该方法通过研究中文组块分析任务及其序列化特性,采用条件随机域模型融合不同类型特征,克服标记偏置问题,将语义词典中抽取的语义类特征应用到中文组块分析中,提高分析精度. 实验表明,该方法取得了 F 值为 92.77% 的中文组块分析性能,实验进一步还表明了特征模板的选取和训练语料的规模对于分析性能的影响.

关键词: 条件随机域;中文组块分析;特征模板;语义词典
中图分类号: TP391.2 **文献标志码:** A **文章编号:** 0367 - 6234(2011)07 - 0135 - 05

Chinese chunking method based on conditional random fields and semantic classes

SUN Guang-lu[1,2], LANG Fei[3], XUE Yi-bo[1]

(1. Research Institute of Information Technology, Tsinghua University, 100084 Beijing, China, guanglu. sun@ gmial. com;
2. School of Computer Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China;
3. School of Foreign Languages, Harbin University of Science and Technology, 150080 Harbin, China)

Abstract: To improve the accuracy of Chinese chunking and utilize the semantic information of words, a new Chinese chunking method is proposed based on conditional random fields and semantic classes. Through the analysis of Chinese chunking task and its sequential characteristics, conditional random fields that could incorporate various types of features were applied to overcome the label bias problem. Semantic features were utilized to improve the chunking performance. Experimental results show that the algorithm achieves impressive accuracy of 92.77% in terms of the F-score. A further experiment indicates the effects of feature template selection and training data's scales on the aspect of chunking performance.

Key words: conditional random fields; Chinese chunking; feature template; semantic dictionary

自然语言处理让计算机能够对人类语言进行处理和结构化,乃至完全理解人类语言. 其包含一系列关键技术,以字、词、短语、句子、篇章的顺序逐层地对自然语句进行标记、分析和处理. 这些技术主要有:字处理技术、分词、词性标注、命名实体识别、组块分析、完全句法分析和语义分析等.

收稿日期: 2008 - 01 - 23.
基金项目: 国家自然科学基金资助项目(60903083);黑龙江省自然科学基金项目(F200936);黑龙江省高等学校新世纪优秀人才基金资助项目(1155-ncet-008).
作者简介: 孙广路(1979—),男,副教授,硕士生导师;
薛一波(1967—),男,研究员,博士生导师.

组块分析(chunking),也称作部分句法分析(partial parsing)或浅层句法分析(shallow parsing),由 Abney 提出[1]. 它以句子的词法信息(包括分词标记和词性标记)为基础,对句子进行句法级的标记. 具有良好性能的组块分析系统可以提供自然语句的浅层句法信息,满足很多语言信息处理系统的需求,同时对更深层的语言分析技术提供有力的支持. 本文提出了基于条件随机域模型的中文组块分析算法,在开放测试中的性能优于基于最大熵马尔科夫模型的组块分析方法;在条件随机域模型中引入了语义类特征,进一步提升了分析性能.

1 中文组块分析的研究现状

现有中文组块分析的研究主要包含 2 个方面的内容:1)中文组块的定义及语料库的生成;2)分析算法的研究与实现.

参照 CoNLL-2000 会议对于英文组块的定义[2],一些学者建立了中文组块的定义和相应语料库,具有代表性的主要有 2 类:1)沿用中文句法树库中的句法标记和短语划分,确定若干组块抽取规则,在句法树库中直接抽取非终结节点作为组块[3];2)基于语言学家对于组块的定义和具体语言现象的分析,将中文文档进行人工标注组块标记,构造独立的中文组块定义及语料库[4-5].相对于第 1 类,第 2 类定义方法不与句法树相关,不需要定义抽取规则和保持抽取一致性,更具有组块定义的独立性和完整性.

组块分析的算法主要包括 3 类:1)基于规则的方法,如文献[1]提出的基于有限状态自动机的方法、文献[6]提出的基于错误驱动的规则匹配方法;2)基于统计的方法,如文献[3]提出的基于最大熵模型的方法、文献[4]提出的基于最大熵马尔科夫模型的方法,文献[7]提出的基于支持向量机的方法,文献[8]提出的基于条件随机域的方法等;3)规则和统计相结合的方法,如文献[9]提出的手工规则和基于实例学习算法相结合的方法. 后 2 类方法是当前研究的主流方法,它们都是在统计方法的基础上,试图融合更多的具有描述能力的特征,用以提升组块分析的性能. 在上述基于不同模型的算法中,所采用的特征主要包含词特征、词性特征以及词缀特征.

本文采用了微软亚洲研究院(MSRA)建立的组块定义、标记集和语料库. MSRA 中文组块分析语料库是专门面向组块分析问题定义和标注的. 语料库包含了人工标注组块标记的近 50 万词中文新闻语料,通过对自然语言现象的分析,有效地解决了组块定义的不一致性和复杂结构的歧义性等问题,为中文组块分析提供了坚实的基础.

条件随机域模型是由 Lafferty[10] 提出的有指导的机器学习模型. 该模型在观测序列的条件下对标记序列进行建模,是一种典型的条件概率模型,重点解决序列化标注的问题. 条件随机域模型既具有条件概率模型的直接对标记问题建模,不需要很强的独立性假设,可以融合多种特征的特点;又具有生成模型(如隐马尔科夫模型)的考虑到标记间的转移概率,以序列化的形式进行参数优化和解码的特点,解决了其他条件概率模型(如最大熵马尔科夫模型)难以避免的标记偏置问题. 由于条件随机域模型具有上述特点,而且中文组块分析问题可以被转化成基于标记间转移的序列化标注问题,故而其适于解决中文组块分析的问题.

对于模型中应用特征的选取,在选取词特征、词性特征和词缀特征的基础上,还通过对于语义词典《同义词词林(扩展版)》的抽取,定义了语义类特征来帮助提升中文组块分析的性能. 语义词典根据词的语义特征为词定义了不同的语义类,其对于句子的组块分析有 2 点帮助:1)利用词典中词的规模及分类类别解决训练语料库中一部份词的数据稀疏问题;2)词的语义信息标定可以帮助提升组块分析的性能[11].

2 中文组块的定义和类型

组块是一个被标记了句法功能标记的非递归、非嵌套、不重叠的词序列. 英文组块内部一般包含一个中心成分以及中心成分的前置修饰成分,而不包含后置附属结构. 组块严格按照句法形式定义,而不体现语义性或者功能性. 本文定义的中文组块需要遵循 2 条基本原则:

1) 组块不能够破坏句子固有的短语结构,主谓结构和动宾结构不能出现在一个组块中.

2) 组块具有一种平整的结构,不再划分组块内部的词之间以及组块之间的关系.

MSRA 中文组块语料库是在北京大学开发的"1998 年 1 月人民日报分词和词性标注公开语料库"[12]的基础上,人工进行组块标记而成的. 语料库包含了 42 种词性标记和 11 种组块类型标记. 为了便于识别组块的边界,定义了组块的 4 种边界标记. 其中:"*B*" 为开始;"*I*" 为中间;"*E*" 为结束;"*S*" 为单个词组块. 对于一些特殊的助词和连词(例如:"的","和","与","或"),定义了它们不属于任何组块,并用"*O*" 来标记它们. 本文将 11 种类型标记和 4 种边界标记结合在一起,再加上"*O*" 标记,一共定义了 45 种标记单个词的组块标记.

3 基于条件随机域模型的组块分析算法

3.1 条件随机域模型

给定一个观测序列 X,基于条件随机域模型的标记序列 Y 的条件概率为

$$p\left(\frac{Y}{X}\right)=\frac{1}{Z(X)}\exp\left\{\sum_k l_k f_k(y_{i-1},y_i,X,i)\right\}. \quad (1)$$

式中 $Z(X) = \sum_{y \in Y} \exp\{\sum_k l_k f_k(y_{i-1}, y_i, X, i)\}$ 为归一化因子.

$f_k(y_{i-1}, y_i, X, i)$ 是条件随机域中通用的特征定义形式，可以被分解为 2 种具体的特征定义：

1）边特征（转移特征）为

$$t_{y',y}(y_{i-1}, y_i, x_i) = \begin{cases} 1, & \text{if } y_{i-1} = y', y_i = y \text{ and } x_i = x; \\ 0, & \text{else.} \end{cases} \tag{2}$$

2）顶点特征（状态特征）为

$$s_{y,x}(y_i, x_i) = \begin{cases} 1, & \text{if } y_i = y \text{ and } x_i = x; \\ 0, & \text{else.} \end{cases} \tag{3}$$

由式（2）~（3），式（1）可以被分解为

$$p\left(\frac{Y}{X}\right) = \frac{1}{Z(X)} \exp \cdot \{\sum_k (\mu_k t_k(y_{i-1}, y_i, X, i) + \zeta_k s_k(y_i, X, i))\}. \tag{4}$$

式中 μ_k 和 ζ_k 分别为转移特征和状态特征的权重参数. 基于最大似然估计原理和 L-BFGS 算法，模型进行参数训练，使得 $p\left(\frac{Y}{X}\right)$ 的对数似然度最大化为

$$L_\Lambda = \sum_{i=1}^{N} \log\left(P_\Lambda\left(\frac{y_i}{x_i}\right)\right) - \sum_{k=1}^{K} \frac{l_k^2}{2\sigma_k^2}. \tag{5}$$

式中：第 2 项为采用提供平滑处理的特征参数的高斯先验值；σ_k^2 为 k 维特征的方差. 由于 L-BFGS 算法只需要对数似然度的一阶导数，则对式（5）求导为

$$\frac{\partial L_\Lambda}{\partial l_k} = \sum_{i=1}^{N} f_k(y_{i-1}, y_i, X, i) - \sum_{i=1}^{N} P_\Lambda\left(\frac{y_i}{x_i}\right) f_k(y_{i-1}, y_i, X, i) - \frac{l_k}{\sigma_k^2}. \tag{6}$$

最后利用动态规划算法求得最优序列 Y^* 为

$$Y^* = \arg\max_Y \left\{p_l\left(\frac{Y}{X}\right)\right\}. \tag{7}$$

3.2　语义类特征抽取算法

《同义词词林（扩展版）》是一种语义词典，是语言学家根据对于语言的理解和统计知识构造的. 它利用树状结构将收录的词条分为 5 层，各层包含的词汇类别分别为 12、95、1 425、4 223 和 17 807 类. 由于其对于词的语义分类不唯一，在使用中容易造成分类歧义. 本文采用投票机制，根据歧义词的词性特征来消除分类歧义，从而建立语义类特征抽取算法，如图 1 所示.

其中，result[ ] 为查询到的语义类结果. 当查询结果为多个语义类标记时，Fur _ Proc（result）的方法为：

1）根据{POS}，考察 result[i] 中包含的每个词的词性标记，将具有唯一词性标记的词抽出；

2）利用抽出词的词性投票机制得到该语义类体现的主要词性特征；

3）选择与 w_i 词性特征相匹配的语义类标记 w_i. tag；

4）若多个 result[i] 体现了相同的词性特征，利用类中词数平均权值的方法选择最好的语义类标记 w_i. tag；

5）若无语义类的词性特征与当前词相对应，则 w_i. tag = null.

```
输入：
{w_i} //语料库中的所有词.
{Sem}//同义词词林.
{Pos}//词性词典.
算法：
For each word w_i,
  result[ ] = Search (w_i, {Sem});
    switch(sizeof(result))
      {case 1: w_i. tag = result[0];
       case 0: w_i. tag = null;
       default: w_i. tag = Fur_Proc(result);}.
```

图 1　语义类特征抽取方法

3.3　特征选择

条件随机域模型的性能在很大程度上依赖于特征模板的选取. 对于词序列 $W = w_1, w_2, \cdots, w_k$，选取宽度为 5 的窗口，抽取当前词 w_i 和前后各 2 个词的词特征、词性标记特征和语义类特征. 另外，还抽取了 w_i 的前后各 1 个字的词缀特征. 不同的特征模板的组合也会影响系统性能的表现. 系统选取了上述原子特征的 Bigram 组合作为复合特征模板. 此外，由于低频特征掺杂了很多的噪声，其统计特性比较差，而且条件随机域模型训练的时间复杂度很高，选择了原子特征出现次数 >5 和复合特征出现次数 >3 的特征作为实际系统训练中使用的特征.

通过语义类特征抽取算法中的每个词和组块标记的映射关系，将中文组块分析转化为序列化分析和标记的任务来进行处理. 给定由词序列 $W = w_1, w_2, \cdots, w_k$ 组成的句子，其相应的词性序列为 $P = p_1, p_2, \cdots, p_k$，语义类标记为 $S = s_1, s_2, \cdots, s_k$，句子可以被划分成若干个组块，每个词 w_i 被标记了组块标记 t_i，$T = t_1, t_2, \cdots, t_k$ 代表组块标记序列. 组块分析的结果为

标记样例：

$$T = \begin{matrix} \cdots [w_i/p_i/s_i \cdots w_{i+m}/p_{i+m}/s_{i+m}] & [w_{i+m+l}/p_{i+m+l}/s_{i+m+l} \cdots w_{i+m+h}/p_{i+m+h}/s_{i+m+h}] \cdots \\ \cdots [t_i \quad \cdots t_{i+m}] & [t_{i+m+l} \quad \cdots t_{i+m+h}] \cdots \end{matrix}.$$

4 实验结果和分析

4.1 MSRA 中文组块分析语料库

语料库包含 19 239 个句子,257 860 个中文组块和 501 804 个词,从语料库抽取出的词典包含 34 830 个词,包含 42 种词性标记和 11 种组块类型标记. 组块的平均长度为 1.507 个词. 本文将语料库分为训练集、开发集和测试集,表 1 列举了语料库中随机选取的训练集和测试集的统计结果.

表 1 MSRA 中文组块分析语料库统计

语料库	句子数量	组块数量	词数
训练集	17 253	229 989	444 777
开发集	1 000	13 992	28 645
测试集	986	13 879	28 382

4.2 算法实验结果及分析

本文采用通用的性能指标:精确率(P)、召回率(R) 和调和平均值(F) 来评价组块分析的性能. 所有的实验都是在开放测试的条件下进行的. 为了验证基于条件随机域模型的算法的正确性和性能,选择 CoNLL-2000 任务中的英文组块公开语料进行了实验,并和性能最好的基于 SVM 的结果进行比较,得到了可比的实验结果,如表 2 所示.

表 2 CoNLL-2000 英文组块分析语料的实验结果

语料库	基于 SVM 的算法	本文算法
CoNLL-2000	93.48	93.32

在公开语料库上验证了算法正确的前提下,为了与已有的工作相比较,首先只应用了词、词性和词缀特征进行实验. 且将训练语料分成 10 等份,采用逐份增加训练语料的方法,发现组块分析性能随着训练语料的增加有不同程度的提升. 图 2表明了在相同特征和语料库规模的条件下,基于条件随机域模型的算法取得的性能都优于基于最大熵马尔科夫模型的算法. 在使用全部训练语料的情况下,基于条件随机域模型的组块分析算法的性能 F 值是 92.0%,比基于最大熵马尔科夫模型的算法的最优性能 91.02% 的 F 值提升了约 1%. 这说明条件随机域模型在克服了最大熵马尔科夫模型所具有的标记偏置问题后,在解决序列化标注问题时表现出了非常好的性能. 同时还发现,将训练语料增加至全部训练语料的 1/2 以上时,性能曲线趋向于平缓,提升不再明显. 这说明在组块分析算法达到较高性能后,增加训练语料对性能帮助有限.

在加入语义类特征的情况下,分别利用不同层次的语义类特征进行组块分析实验,在利用开发集数据调优后,发现选取 4 层语义类可以得到最优结果.

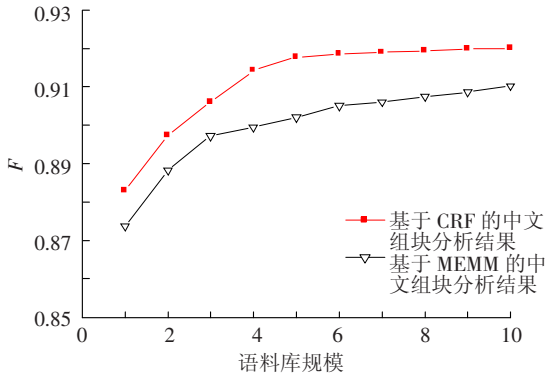

图 2 基于条件随机域模型和最大熵马尔科夫模型的组块分析算法性能比较

表 3 分别列举出了基于第 4 层语义类特征的组块分析算法对每一类中文组块进行标记的结果,并统计了每一类组块的平均长度(以中文词为基本单位),以及在语料库中所占的比例. 从这些数据中可以看出,名词性组块和动词性组块占语料库的 75% 以上,它们是中文句子的主要组成部分,对它们的识别精度很大程度上决定了组块分析的整体性能. 算法对于平均长度最长的插说组块的分析性能非常好,这也证明了条件随机域模型可以较充分地考虑到上、下文特征,并对整个序列进行参数寻优,对于上、下文结合紧密的序列的分析取得了较好的效果. 在加入了语义类特征后,整体性能有了明显的提升,几乎每类组块分析结果也都有了不同程度的提升.

表 3 基于第 4 层语义类的中文组块分析性能

标记类别	平均长度	语料库中比例/%	F/ %	F'/ %
名词性组块	1.649	45.94	89.80	88.79
动词性组块	1.416	29.82	97.16	96.83
介词组块	1.221	6.59	95.87	95.98
数词性组块	1.818	3.69	90.07	88.78
形容词性组块	1.308	3.77	90.01	87.57
方位组块	1.167	2.71	84.18	82.88
时间组块	1.251	2.59	93.89	94.12
连词组块	1.000	2.22	97.97	97.97
插说组块	4.297	1.41	98.59	97.20
副词性组块	1.117	1.06	90.14	87.01
语气组块	1.016	0.23	95.65	97.87
总计	1.507	100	92.77	92.00

注:F' 为不利用语义类特征模型的 F 值结果.

图 3 中的 4 条曲线分别代表只采用词特征的模板、只采用词性特征的模板、采用词、词性和词

缀特征模板以及包含所有特征的模板的组块分析算法性能. 本文同样采用逐渐增加语料库规模的方式来分析性能的变化. 从图 3 中可以看出，只采用词性特征的算法性能明显优于只采用词特征的算法. 由于词典词规模为 3 万多，而词性规模只有 42 个，数据稀疏的问题导致了词特征模型的性能不佳，但是从其性能曲线可以看出，增加语料库规模仍然可以大大提升词特征模型的性能. 词性特征对于组块分析有着较强的预测能力，但是利用该特征的模型在训练语料规模达到 1/2 时，性能已经达到最优值. 综合了词和词性特征的模型取得了较优的性能，而增加了语义类特征后的算法性能曲线在不同语料规模下都达到最优值，且性能曲线随着语料库规模的增大还在缓慢而持续的上升. 增加了语义类特征后，即使使用 1/10 规模的训练语料，分析性能也能达到 90% 以上，证明在语义类的帮助下，较小规模的训练语料也可以达到很好的性能. 由此可见，词性特征和语义类特征对于组块分析起到了类别知识和数据平滑的作用，对于组块分析的性能有较强的指示作用；而词特征对于组块分析起到了判别和实例化的作用.

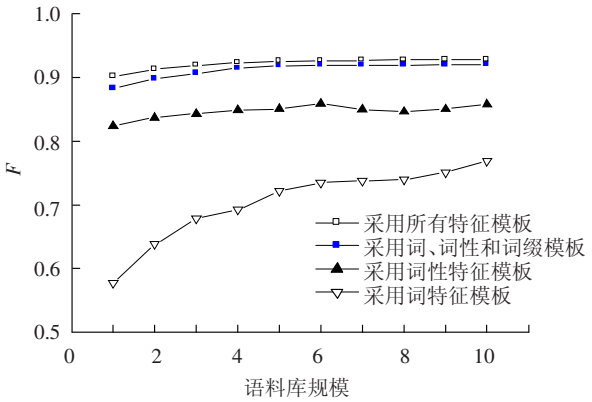

图 3　各种特征集和训练语料规模下基于条件随机域模型的组块分析算法性能

5　结　论

1）分析了条件随机域模型在序列化标记任务中的优势，将其应用到中文组块分析中，结合 MSRA 中文组块语料库，取得了 F 值为 92% 的分析性能，比基于最大熵马尔科夫模型的分析算法提升了约 1%.

2）利用语义词典抽取语义类特征，将其加入分析模型，算法性能进一步提升，得到 92.77% 的最优性能.

3）研究了不同类型特征对于组块分析性能的影响和对于训练语料规模的需求. 词性特征和语义类特征对于组块分析有着较强的预测能力，其对于语料库的规模需求较小；词特征对于分析有着判别和实例化的作用，结合其他特征共同使用可以进一步提升系统性能.

参考文献：

[1] ABNEY S, ABNEY S P. Parsing by Chunks: Principle-Based Parsing[M]. Dordrecht: Kluwer Academic Publishers, 1991: 257 – 278.

[2] TJONG KIM SANG E, BUCHHOLZ S. Introduction to the CoNLL-2000 shared task: Chunking[C]//Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning. Stroudsburg, PA: Association for Computational Linguistics, 2000: 127 – 132.

[3] 李素建，刘群，杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报，2003，25(12)：1722 – 1727.

[4] SUN G, HUANG C, WANG X, *et al.* Chinese chunking based on maximum entropy markov models[J]. International Journal of Computational Linguistics and Chinese Language Processing, 2006, 11(2): 115 – 136.

[5] 周强，李玉梅. 汉语块分析评测任务设计[J]. 中文信息学报，2010，24(1)：123 – 128.

[6] RAMSHAW L A, MARCUS M P. Text chunking using transformation-based learning[C]//Proceedings of the 3rd ACL/SIGDAT Workshop. Cambridge, MA: Association for Computational Linguistics, 1995: 222 – 226.

[7] 周俊生，戴新宇，陈家骏，等. 基于大间隔方法的汉语组块分析[J]. 软件学报，2009，20(4)：870 – 877.

[8] CHEN Wenliang, ZHANG Yujie, ISAHARA H. An empirical study of Chinese chunking[C]//Proceedings of the Coling/ACL on Conference Poster Sessions. Stroudsburg, PA: Association for Computational Linguistics, 2006: 97 – 104.

[9] PARK S B, ZHANG B T. Text chunking by combining hand-crafted rules and memory-based learning[C]//Proceedings of the 41st Annual Meeting of ACL. Stroudsburg, PA: Association for Computational Linguistics, 2003: 497 – 504.

[10] LAFFERTY J, McCALLUM A. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc, 2001: 282 – 289.

[11] XIONG Deyi, LI S, LIU Q, *et al.* Parsing the penn chinese treebank with semantic knowledge[C]//Proceedings of IJCNLP – 2005. Berlin: Lecture Notes in Computer Science, 2005: 70 – 81.

[12] YU S, DUAN H, ZHU X, *et al.* The basic processing of contemporary chinese corpus at peking university[J]. Journal of Chinese Information Processing, 2002, 16(6): 58 – 65.

（编辑　张　红）