基于条件随机场的有标记联合结构自动识别

王东波，陈小荷，年洪东
（南京师范大学文学院，210097）

摘要：本文详细介绍了 CRF 基本原理，并用该模型对有标记联合结构进行了自动识别。分别用基于复杂特征的特征模板和增加语言学特征的特征模板在含有嵌套的联合结构、无嵌套联合结构和最长联合结构语料上进行了实验，封闭测试和开放测试调和平均值最高分别达到：99.17%和 88.21%；99.99%和 87.85%；99.98%和 84.42%。

关键词：有标记联合结构；条件随机场；特征模板；

The Automatic Identification of Coordination with Overt Conjunctions, Based on CRF

Wang Dongbo,Chen Xiaohe,Nian Hongdong
（School of Chinese Language and Literature, Nanjing Normal University，210097）

Abstract： The article introduces the basic principle of CRF(Conditional Random Fields) and uses it to identify the COC. The article has respective test in the corpora which includes nesting COC, non-nesting COC and longest COC by the model of complicated features and model with the linguistic features. The best F scale of recall and precision in the COC respectively reaches 99.17% and 88.21%,99.99% and 87.85% and 99.98% and 84.42% in the closed and open tests.

Keywords: Coordination with Overt Conjunctions; Conditional Random Fields ; Model of Feature;

1 引言

从整体上说联合结构分为无标记联合结构和有标记联合结构。有显性标记形式的联合结构称之为有标记联合结构，如“建设/vn 方案/n 、/w 项目/n 调整/vn 和/c 概算/n”和“选育/vn 与/c 推广/vn ”，其中“、”与“和”就是联合标记。有标记联合结构分布广泛、结构复杂、长度跨度大，导致了该结构识别起来十分困难。周强在《汉语语料库的短语自动划分和标注研究》博士论文中，把联合结构放在整个句子生成的层面验证，得出了联合结构的识别“错误很严重”的结论。孙宏林在《现代汉语非受限文本的实语块分析》博士论文中也谈到了联合结构的处理。利用并列成分之间的对称性，通过一个简单的概率模型来识别联合结构的边界，但效果也不是很理想。本文在前人研究的基础上，简单地探讨了使用 CRF 模型来识别有标记联合结构，得到了相对满意的结果。

2 CRF 模型概述

条件随机场是一个无向图模型的框架，它能够被用来定义在给定一组需要标记的观察序列的条件下，一个标签序列的联合概率分布。假定 X 是将要被标注的数据序列的随机变量，Y 是相应的标签序列的随机变量。例如，X 是自然语言的句子集合，Y 是标注这些句子的词性集合。随机变量 X 和 Y 是联合分布的，根据观测序列和标签序列对，构建了一个条件模型 p(Y|X)。

定义：图 G=(V，E)是一个无向图，如果给定 X，随机变量 $Y=(Y_v)_{v\in V}$ 遵循马尔可夫属性，即 $p(Y_v \mid X, Y_w, v\neq w)=p(Y_v \mid X, Y_w, v\sim w)$，$v\sim w$ 表示 v 和 w 是 G 中相邻的节点，那么(X，Y)是条件随机场。

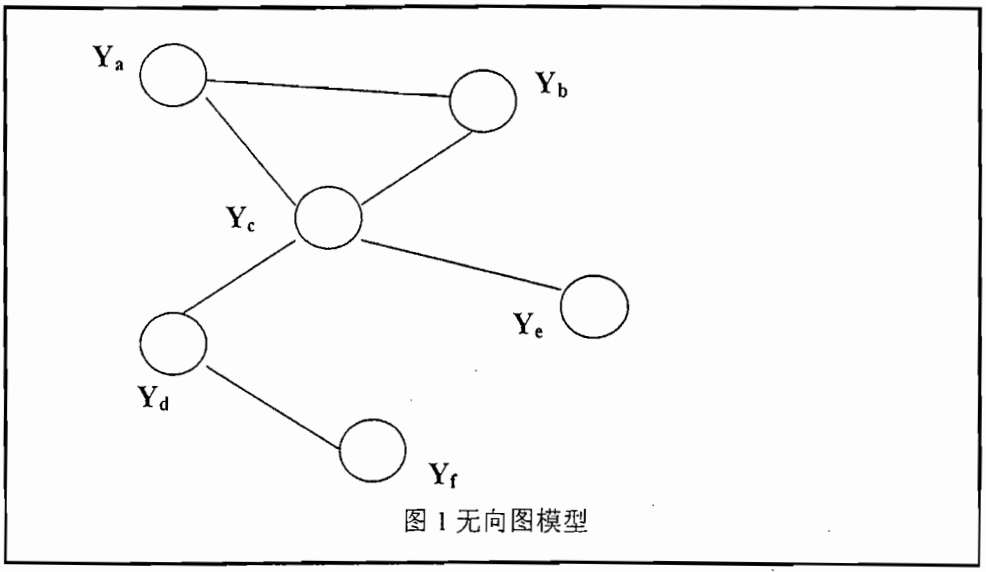

图 1 无向图模型

给定观测序列 $X=(X_1,X_2,\ldots\ldots,X_n)$，标签序列 $Y=(Y_1,Y_2,\ldots\ldots,Y_n)$的情况下，在图 G=(V,E)中，Y 是一棵树(最简单的情况下是一个链结构)。因此，根据随机场的基础理论(Hammersley & Clifford,1971),X 和标签序列 Y 的联合概率的形式是：

$$P_\theta(y \mid x) \propto \exp\Big(\sum_{e\in E,k}\lambda_k f_k(e, y|_e, x)+\sum_{v\in V,k}\mu_k g_k(v, y|_v, x)\Big)$$

其中 x 是数据序列，y 是标签序列，v 是顶点集合，e 是边的集合，k 是特征数。例如，词 X_i 是大写的，标签 Y_i 是“proper noun”，那么布尔特征 g_k 可能为真。

联合结构自动识别的任务可以看成文本中词语与词性序列选择标记、确定边界的过程，可以利用 CRF 来确定边界，但由于有标记联合结构的长度比较复杂，因此标记不能是非此即彼的两个标记，这个标记序列应该有一些标记构成。

3 CRF 识别有标记联合结构语料的预处理

本文中所有有关CRF的相关实验都是使用基于C++语言开发的CRF++工具进行的，可以

在网站下载并使用，网址为：http://chasen.org/~taku/software/CRF++/#features。

在本文的实验中，我们分别使用了北大1998年1月份前十天56万字的语料和清华一百万的语料作为训练和测试的依据，这些语料都是经过了分词和词性标注过的熟语料并且把文本中的有标记联合结构用｛ ｝和[ ]表示出来了。由于北大人民日报语料和清华语料分词和词性标记用的不是一个标准，所以就没有把这两个语料库合在一起训练和测试，而是进行了分别训练和测试，在一定程度上方便进行作对比实验。

本文在确定用于联合结构识别的CRF标记数的时候，参考了下面这个公式，

$$L_k = \frac{1}{N}\sum_{i=k}^{k} iN_k \quad (k>2)$$

其中，L_k是 i>=k 时的平均加权有标记联合结构长度，N_k是语料中联合结构长度为 k 出现的次数，k 是语料中出现过的最大联合结构长度，N 是语料库中联合结构总的出现次数。如果 k=2，那么 L_2代表整个语料的联合结构的平均长度。在参考这个公式的前提下，根据具体的实验结果，本文确定使用 7 词位标注集，具体的标注集为 T={B，F，G，I，M，E，S}，其中 B 是有标记联合结构的开始词，F，G，I，M 是结构中的词，E 是结构结尾的词，S 是有标记联合结构外部的词。

4 联合结构识别中CRF特征的选取以及特征模板的确定

特征是基于CRF的联合结构自动识别的核心，特征选择的好坏将影响CRF模型识别的性能。根据是否加入特殊的语言学特征，联合结构自动识别的分为两类，一类基于复杂特征的联合结构识别，另一类基于增加语言学特征的联合结构识别。

基于复杂特征的联合结构自动中的复杂特征主要是由词语和词性产生出来的，词语和词性用标记表示为“W，P”，字母旁边的整数表示所考察的特征位置。例如 0 表示当前位置、－1 表示左边第一个位置、1 表示右边第一个位置。我们所选择的特征窗口分别为：词语 7 个窗口，范围是{-3,-2,-1,0,1,2,3}，词性为 5 个窗口，范围是{-2,-1,0,1,2}。根据观察窗口，结合联合结构识别确定特征模板的实验，基于复杂特征联合结构识别共使用了 18 个特征，具体为：

W-3,W-2,W-1,W,W+1,W+2,W+3,W-1/W,W/W+1,W-1/W+1,P-2,P-1,P,P+1,P+2,P-1/P,P/P+1,W/P;

CRF 本身的一个突出优点是可以任意加入与处理的对象有关的语言学特征，作为一个独立的语言学结构，联合结构有其自身的特征。本文在用 CRF 识别联合结构时，增加的语言学特征具体有下面这些：

词语的长度、词语的拼音、词语是否是连词、词语是否是边界词.

在复杂特征模板的基础上，增加上面这四个语言学特征后，构成了新的 CRF 特征模板。在新模板的基础上，结合 7 词位标注集形成的新的训练语料如表 1 所示：

表 1 语言学特征模板下的训练语料样例

词语	词性	词语长度	是否连词	是否边界词	词语拼音	标记
同胞	n	2	N	N	ong2bao1	B
们	k	1	N	N	men5	F

、	w	1	N	Y	wu2	G
朋友	n	2	N	N	peng2you5	I
们	k	1	N	N	men5	M
、	w	1	N	Y	wu2	M
女士	n	2	N	N	nv3shi4	M
们	k	1	N	N	men5	M
、	w	1	N	Y	wu2	M
先生	n	2	N	N	xian1sheng5	M
们	k	1	N	N	men5	E
：	w	1	N	N	wu2	S
在	p	1	Y	N	zai4	S
１９９８年	t	5	N	N	wu2	S
来临	v	2	N	N	lai2lin2	S
之际	f	2	N	N	zhi1ji4	S

在表 1 中，凡是标点符号，长度都设定为 1，其中数字或外语字符串都是按照单个字符计算长度的，如“1998”长度就设定为 4。给词语加注拼音的时候，由于 CRF 训练模型不允许某一列存在空缺的特征标注，所以像标点符号、没有在拼音词典中出现的词语都被标注成了“wu2”的形式。

有标记联合结构自动识别的框架结构如图 2 所示

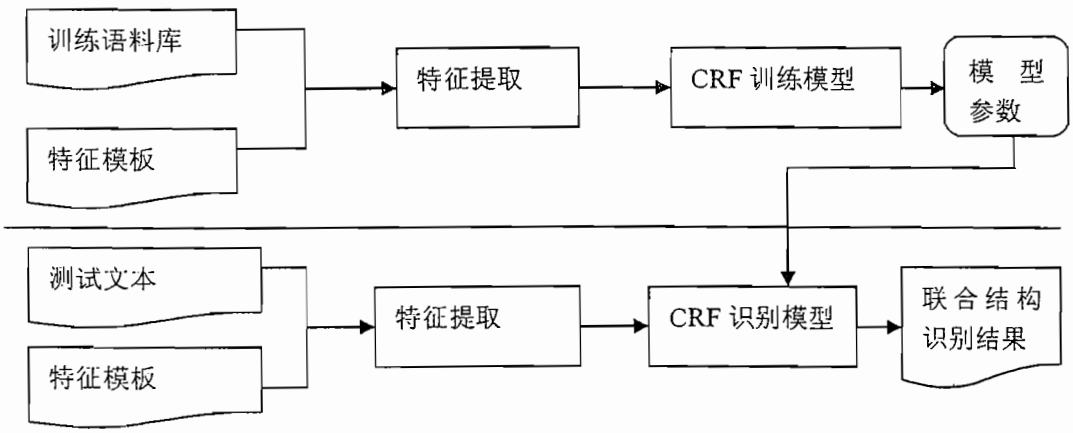

图 2 基于 CRF 的有标记联合结构识别系统的框架结构

5 实验设计和实验结果分析

本文的实验在北大语料和清华语料的基础上基于检验特征模板的有效性和方便两个语料识别效果对比的目的，具体设计了三个实验并分析了实验结果。

5.1 含有嵌套联合结构的实验

本实验所用训练语料规模为：北大人民日报语料和清华语料各 50 万左右，根据封闭测试和开放测试 9:1 的语料比例，利用随机抽样的方法，从两个语料库中选取了各 6 万左右的语料作为开放测试用。语料中联合结构内部嵌套的联合结构仍然保留，仍然按照标注集进行标注，具体见附录。封闭测试和开放测试的结果仍然用精确率、召回率和调和平均值来衡量。

在这两个语料规模上，分别使用特征模板一和特征模板二获得 CRF 训练模型，前一个特征模板没有增加语言学特征，后一个则包含了语言学特征。具体开封闭测试和开放测试结果如表 2、3、4、5 所示

表 2 基于模板一的含有嵌套联合结构封闭测试实验结果

测试语料	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
北大语料	5843	99	3	98.33%	99.95%	99.13%
清华语料	5091	138	3	97.36%	99.94%	98.63%

表 3 基于模板一的含有嵌套联合结构开放测试实验结果

测试语料	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
北大语料	532	127	20	80.73%	96.38%	87.86%
清华语料	432	138	22	76.73%	95.15%	84.95%

表 4 基于模板二的含有嵌套联合结构封闭测试实验结果

测试语料	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
北大语料	5098	83	2	98.40%	99.96%	99.17%
清华语料	5091	141	0	97.31%	100.00%	98.63%

表 5 基于模板二的含有嵌套联合结构开放测试实验结果

测试语料	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
北大语料	550	130	17	80.88%	97.00%	88.21%
清华语料	448	112	23	80.00%	95.12%	86.91%

从表 2 和表 4 可以看出,CRF 模型的封闭测试效果非常好，在具有语言学特征的模板二中，整个调和平均值达到了最高值，为 99.17%。

从表 3 和表 5 中可以看出，虽然整个开放测试的调和平均值都在 84.95%以上，但无论是用模板一还是用模板二来识别联合结构，精确率都非常差，其中在模板一中，清华语料的精确率仅为 76.73%。造成这种结果的主要原因是嵌套的联合结构比较复杂，在识别的时候容易识别错误，而考虑嵌套的联合结构后，增加了联合结构标记的数量，召回率就有所提高。

从两个模板识别的结果对比来看，增加了语言学特征的模板二的识别效果在开放测试和封闭测试上都取得了比基于模板一要好的效果，但这种效果是非常小的，从调和平均值上看仅仅增加了一两个百分点。

5.2 无嵌套联合结构的实验

根据统计有标记联合结构内部出现联合结构的比率非常低，因此从整体上把有嵌套的联合结构去掉进行实验，对整个结果影响不会太大。语料仍然是 50 万左右的北大和清华的语料，根据 9:1 的比例用随机抽取的软件把语料分成训练和测试两部分，由于去掉了带嵌套的联合结构，所以这次封闭测试和开放测试里面的联合结构与含有嵌套联合结构的实验里的联合结构是不完全一样的。具体的封闭测试和开放测试结果如表 6、表 7、表 8 和表 9 所示。

表 6 基于模板一的无嵌套联合结构封闭测试实验结果

测试语料	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
北大语料	5483	2	0	99.936%	100.00%	99.968%
清华语料	6098	0	4	100.00%	99.93%	99.965%

表 7 基于模板二的无嵌套联合结构封闭测试实验结果

测试语料	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
北大语料	5455	19	3	99.65%	99.95%	99.97%
清华语料	6100	0	2	100.00%	99.97%	99.99%

表 8 基于模板一的无嵌套联合结构开放测试实验结果

测试语料	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
北大语料	465	114	22	80.72%	95.48%	87.48%
清华语料	504	67	125	88.27%	80.13%	84.00%

表 9 基于模板二的无嵌套联合结构开放测试实验结果

测试语料	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
北大语料	470	114	16	80.48%	96.71%	87.85%
清华语料	512	69	127	88.12%	80.13%	83.94%

两个模板的封闭测试效果要好于有嵌套的联合结构识别效果，最差的调和平均值为 99.965%，从中可以看出 CRF 模型在数据拟和性上具有非常突出的优势。

整体开放测试的效果稍微要比有嵌套的联合结构识别差些，主要原因是把嵌套联合结构抽取后破坏了整个联合结构的上下文语境，造成了训练模型在统计的时候知识中断。

具有语言学知识的特征模板识别效果并不比单一的复杂特征模板好多少，在封闭测试上基本上是一致的，而在开放测试的时候特征模板二的识别效果在北大语料上比特征模板一高 0.37%在清华语料的识别结果上反而低 0.006%。这说明越是简单、明显的特征，其影响效果就越大，比如词性标记特征，而一些复杂的特征加入后，反而会影响整体的识别效果。

5.3 最长联合结构的实验

如果把整个嵌套的联合结构从语料中去除，一方面会造成模型统计知识的缺失，另一方面会把一些非嵌套的联合结构错误的删除掉从而使数据稀疏的问题变的更加严重。所以本文设计了一个把嵌套的联合结构看成一个联合结构的实验，这个联合结构称为最长联合结构。

本节实验所用语料为 100 万清华语料，也是按照 9:1 的比例用随机抽样软件把 100 万语料分成训练和测试两部分。封闭测试和开放测试结果见表 10 所示

表 10 最长联合结构实验结果

测试方式	正确识别	错误识别	没有识别	精确率	召回率	调和平均值
封闭测试	8308	0	3	100.00%	99.96%	99.98%
开放测试	580	114	100	83.57%	85.29%	84.42%

封闭测试的效果与上面两个实验没有什么差别，都取得了非常理想的效果。开放测试的调和平均值虽然没有含有嵌套的联合结构特征模板二的识别效果86.91%高，但考虑到模板三所用特征的简单性，其结果还算合理。最长联合结构识别的精确率和召回率比较接近，这从另一个侧面说明了如果语料规模足够大、数据稀疏问题解决的好，从整体上可以提高联合结构识别的精确率和召回率，而不像表3一样一个偏高一个偏低。

6 结论

本文使用CRF统计模型，分别用基于复杂特征的特征模板和增加语言学特征的特征模板在含有嵌套的联合结构、无嵌套联合结构和最长联合结构语料上进行了实验，取得了相对满意的结果。本文下一步的工作主要为：在提高计算机硬件如内存的前提下，扩大训练语料库的规模，更深入地检验CRF的性能；从已标注的联合结构语料中挖掘新的语言学知识，在CRF模型中添加新的语言学特征，进而观察CRF模型的性能。

参 考 文 献

[1] Kurohashi, S. and Nagao, M. (1994). A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. Computational Linguistics, 20(4).

[2]Peng and A. McCallum. 2004. Accurate Information Extraction from Research Papers using Conditional Random Fields. In Proceedings of Human Language Technology Conference and North AmericanChapter of the Association for Computational Linguistics (HLT-NAACL).

[3] Sha and F. Pereira. 2003. Shallow Parsing with Conditional Random Fields. In Proceedings of Human Language Technology Conference and North AmericanChapter of the Association for Computational Linguistics (HLT-NAACL), pp135-136.

[4] 李双龙、刘群. 基于条件随机场的汉语分词系统[J]. 软件天地, 2006(10-1), pp178-179.

[5] 李素建、刘群和白硕. 统计和规则相结合的汉语组块分析[J]. 计算机研究与发展, 2002(4), pp385-386.

[6] 孙宏林. 现代汉语非受限文本的实语块分析[D]. 北京大学博士学位论文, 2001.

[7] 吴云芳. 面向中文信息处理的现代汉语并列结构研究[D]. 北京大学博士论文, 2003, pp2, pp61-62, pp107, pp126.

[8] 周俊生、陈家骏. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报, 2006(5), pp805.

[9] 周强. 汉语语料库的短语自动划分和标注研究[D]（博士学位论文）. 北京大学, 2002, pp37, pp40.

[10] 周雅倩、郭以昆、黄萱菁、吴立德. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展, 2003(3), pp441.