# 基于双层条件随机场的汉语词性标注方法研究

王艺帆[1]，王希杰[2]

（1. 华中科技大学 电子信息与通信学院，湖北 武汉 430000；2. 安阳师范学院 计算机与信息工程学院，河南 安阳 455000）

[摘要] 针对汉语词性标注中词性类别划分较细、类别较多的问题，提出一种利用双层条件随机场进行汉语词性标注的方法，该方法将汉语词性标注分为两个阶段，每个阶段采用一层条件随机场建模实现。第一阶段底层条件随机场根据上下文产生每个词语的词性粗分结果；第二阶段高层条件随机场将词语及其粗分结果作为上下文特征对每个词语的词性进一步细分，产生最终词性标记。利用 CRF + +0.53 工具包，在国际汉语分词评测 Bakeoff2007（国际汉语分词评测）的 NCC 和 CTB 语料上进行了实验，结果表明该方法可行且可以获得较好的标注结果。

[关键词] 汉语词性标注；双层条件随机场；上下文特征；特征模板；词性粗分结果

[中图分类号] TP391　　[文献标识码] A　　[文章编号] 1671 - 5330（2016）05 - 0087 - 05

DOI:10.16140/j.cnki.1671-5330.2016.05.019

## 0　引言

语言中的一个词在词典中可以兼具多种词性，所谓词性标注就是给实际语言环境中的每一个词分配一个唯一的、正确的词性标记。作为自然语言处理领域的一项基础课题，词性标注不仅是句法、语义及篇章理解的基础，也是自动问答、机器翻译、信息检索等后续应用技术的关键[1]。根据使用技术的不同，大致可以将词性标注分为基于规则的标注方法、基于变换的标注方法[2]及基于统计的标注方法，随着近年统计语言模型的普遍应用，用基于统计语言模型的标注方法来解决汉语词性标注问题已经成为了当前主流方法。目前常用的统计语言模型主要有 N 元语法模型[3]、最大熵模型[4]、SVM[5]、隐马尔科夫模型[6]及条件随机场[7,8]等。综合分析这些文献，利用统计语言模型实现中文词性标注的实质就是一个为“词串序列”中的“词”标注合适的词性标签的序列数据标注问题。

条件随机场[9]（conditional random fields，CRFs）是一个用来对序列数据进行标注的优秀的条件概率模型，是由 Lafferty 等人于 2001 年首次提出的。由于条件随机场可以拟合任意的特征，并且有效解决了其它统计语言模型中的标注偏置问题[10]，因此近年来 CRFs 在自然语言处理的许多领域中得到了成功的应用，如汉语自动分词[11]、组块分析和短语识别[12]、命名实体识别[13]等。在使用条件随机场对序列数据标注进行建模时，上下扮演了所需语言知识提供者的角色[9]，在模型训练过程中，CRFs 根据特征模板扩展出的上下文特征统计得出标注过程中所需的语言知识并对这些语言知识进行量化。由于在词性标注中使用的词性标注集合一般都比较大，分类种类较多，导致利用 CRFs 对上下文进行建模时，将会扩展出数以亿计的上下文特征。如此大规模的上下文特征会使得 CRFs 模型的训练时间太长，甚至可能会导致某些条件随机场工具包（例如，CRF + + 工具包）直接崩溃，使得 CRFs 建模过程不能完成。针对这一问题，本文结合条件随机场建模过程，深入分析了上下文特征的表示方法和上下文特征产生的机理，在此基础上提出了基于双层条

[收稿日期] 2016 - 05 - 15

[基金项目] 国家自然科学基金项目（60663004）；河南省高等学校青年骨干教师项目（2009GGJS - 108）

[作者简介] 王艺帆（1995 - ），男，河南安阳人，主要从事自然语言处理、机器学习等研究。

件随机场的汉语词性标注方法，并在 Bakeoff2007 提供的 CTB 和 NCC 语料上进行了大量对比实验，实验结果表明该方法可行且可以获得较好的词性标注结果。

## 1 基于 CRFs 的汉语词性标注

### 1.1 条件随机场定义

条件随机场是一种判定性无向图模型（discriminative model），它以输入结点的值作为条件来计算输出结点值的概率，在给定标记序列 S 和观察序列 O 的情况下，它通过条件概率 P(s/o) 来预测新输入序列中最可能的标记序列。

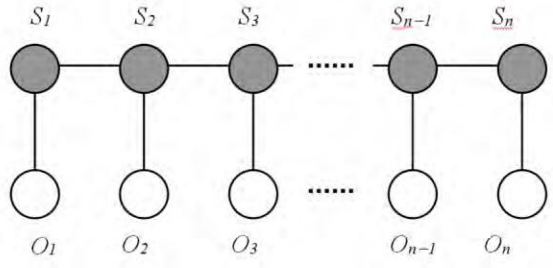


**图 1 线链 CRFs 的结构图**

Fig. 1 structure of linear - chain CRFs

线链条件随机场是一种最简单且最重要的条件随机场，此模型假设在各输出节点间存在一阶马尔可夫独立性，其输出节点被无向边连接成一条线性链（如图 1 所示）。

假设 $O = \{o_1, o_2, \cdots\cdots o_T\}$ 表示输入数据序列（在词性标注中就是待标注词性的“词”序列）；$S = \{s_1, s_2, \cdots\cdots, s_T\}$ 表示被预测的状态序列（对于词性标注问题来说即是词性序列），在给定输入序列的情况下，对于参数为 $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_K\}$ 的 *CRFs* 模型，其输出的状态序列的条件概率为：

$$P_\Delta(S \mid O) = \frac{1}{Z_O}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, o, t)\right) \quad (1)$$

其中，$Z_O$ 是归一化因子，确保所有可能的状态序列的条件概率的和为 1，其定义为：

$$Z_O = \sum_S \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, o, t)\right) \quad (2)$$

$f_k(s_{t-1}, s_t, o, t)$ 为特征函数，通常是一个二值表征函数。

$\lambda_k$ 需要从训练语料中来学习，表示特征函数 $f_k(s_{t-1}, s_t, o, t)$ 的权重，取值范围是 $-\infty$ 到 $+\infty$。特征函数可以整合任何特征，包括状态转移特征 $s_{t-1} \rightarrow s_t$（对词性标注任务，状态转移特征是指词性的转移特征），以及观察序列 $O$ 在时刻 $t$（当前词语位置）的所有特征。

对于由公式（1）定义的 CRFs 模型，在已知输入词串序列 $O$ 的情况下，最可能的词性序列可由公式（3）求出：

$$S^* = \arg\max_S P_\Delta(S \mid O) \quad (3)$$

类似于隐马尔可夫模型，条件随机场模型也可以使用维特比解码（Viterbi decoding）算法来求解最佳的标注序列。

### 1.2 条件随机场对词性标注的建模

使用条件随机场进行汉语词性标注时首先要建立模型，其实质就是用 CRFs 对上下文中的词语序列和词性序列进行建模。采用条件随机场进行汉语词性标注建模时，上下文是正确标注词性所需语言知识的提供者，通常用“上下文特征”来表示这些蕴含于上下文中的特定的语言知识，并用“特征模板”对上下文特征进行描述和表示。可以将特征模板看作是对一组具有共同属性的上下文特征的抽象，其主要功能是定义上下文中某些特定位置的语言成分与某类待预测事件的关联情况。在汉语词性标注中一个词的词性是由该词及其上下文信息来确定的，因此特征模板中应该包含待标注词及该词前、后出现的词或词语组合以及这些词或词语组合出现的位置，表 1 中给出了一些词性标注中常用特征模板及其含义。

**表 1 特征模板列表** Table 1 List of feature templates

| 特征模板 | 模板表征的意义 | 模板扩展出的上下文特征实例 |
|---|---|---|
| $W_{-1}$ | 当前词的前一个词 | 政府 |
| $W_0$ | 当前词 | 顺利 |
| $W_1$ | 当前词的后一个词 | 恢复 |
| $W_{-1}W_0$ | 当前词前一个词和当前词的词语组合 | 政府顺利 |
| $W_0W_1$ | 当前词及其后一个词组的词语组合 | 顺利恢复 |
| $W_{-1}W_1$ | 当前词的前一个词和后一个词的组合 | 政府恢复 |
| $T_{-1}T_0$ | 相邻两个词间的词性转移 | n ad |

表 1 中的 $W_n$ 是单词语特征模板，其中 n 表示某个词和当前词之间的相对距离，如 $W_0$ 表示的就是当前词，$W_1$ 代表当前词后边的一个词，$W_{-1}$ 代表当前词的前一个词，依此类推。除了单词语特征模板，常用的还有双词语特征模板，即每个特征模板由上下文中的两个词语组合而成，表 1 中的前 3 个模板是单词语特征模板，后三个就是双词语特征模板。表 1 中最后边的特征模板 $T_{-1}T_0$ 表示上下文中相邻两个词的词性转移特征。对于“中国/ns 政府/n 顺利/ad 恢复/v ”这个词语序列，在当前词为“顺利”时，各特征模板扩展出的上下文特征如表 1 第三列所示。

在统计语言建模中根据设定的特征模板集可以从训练语料中扩展出大量的上下文特征，如果一个训练语料中包含 3 万个不同的词语，则每个单词语特征模板都会产生 3 万个上下文特征，而双词语特征模板理论上将会产生 9 亿个上下文特征。实质上，Crf + + 工具包产生的上下文特征个数等于上下文特征数乘以给定词性标记集中的标记种类的个数。由于在词性标注中一般所用到的词性标注集合都比较大，因此用 CRFs 对中文词性标注进行建模时，往往会产生数以亿计的特征，这导致模型训练时间过长，甚至无法训练。

### 1.3　基于双层条件随机场的汉语词性标注

在已有的汉语词性标注研究中，语料中所给出的词性标注集合一般都比较细、比较大，如在美国科罗拉多大学（CTB）提供的训练语料中共出现了 32 种词性标记，而中国教育部国家语委（NCC）提供的语料中的词性标记更是达到了 47 种。标注集合之所以这么大，观察其标注集可知，主要是由于对一些词性的划分过细造成的，如在 CTB 语料中仅名词就被分成了 NN、NR、NT 三种类型，而 NCC 语料中的名词更是被细分成了 11 种类型。由前面的分析可知，CRFs 在训练过程中产生的特征数是上下文特征个数与词性标注集中标记种类的个数的乘积，标注集合越大产生的特征就越多，CRFs 在训练的时候就越困难，甚至会出现无法训练的现象。针对这一问题，本文设计了基于双层条件随机场的词性标注方法，其核心思想是将词性标注分为两个阶段进行，第一阶段底层条件随机场按名词、动词等词性产生粗分结果，第二阶段高层条件随机场结合底层的粗分结果，将每一个词的词性进一步细分。第一阶段作者使用自己设计的工具软件对 CTB 和 NCC 语料进行预处理，训练语料的格式分为两列，第一列为词，第二列为词性标记，但这里的词性标记仅仅是粗分结果，即仅仅是名称（n）、动词（v）等词性，并没有对这些词性进行细分。经过第一次预处理后，CTB 语料中的词性标记被减少成了 15 种，NCC 语料中的标记减少到了 21 种，训练语料中的词性标记集合大大减小，使得训练可以顺利进行。训练完成后，使用得到的模型对测试语料进行预测，可以得到每一个词的粗分结果，即每一个词将被标记上名词、动词、形容词等词性。第二阶段的训练语料被分成了三列，第一列是词，第二列是粗分词性，第三列是细分词性，主要目的是让 CRFs 学习每一类词性被进一步细分的概率，如某一个名词被进一步细分成 NN、NR 或是 NT 的概率。在训练完成得到模型后，将第一次预测的带有粗分结果的测试文件作为第二次预测的输入文件，经过预测后得到详细的细分结果，然后将细分结果进行处理，去除掉中间第二列的粗分结果后，和标准答案进行对比评测，最后给出评测结果。

## 2　实验设计及结果分析

### 2.1　实验设计

如前所述，对词性标注建模时，上下文将为条件随机场提供所需的语言知识和相关资源，这就是上下文特征，统计语言建模中用特征模板来表示上下文特征。常用的特征模板有单词性特征模板和双词性特征模板两类，文献[14]对常用的这两类模板进行了定量分析并得出了“双词语特征模板对词性标注的精度没有提高的结论”，根据这一结论，本实验中没有采用双词语特征模板。一个具体的特征模板在模板集中表示的时候，通常以%x [m,n] 来表示，其中的 m 和 n 为整数，m 表示的是以当前字为基准的一个相对行坐标，而 n 表示的是一个以 0 为起始的绝对列坐标。本实验采用的所有特征模板如表 2 所示。

**表 2　特征模板列表**

Table 2　List of feature templates

| 序号 | 特征模板 | 模板扩展出的上下文特征实例 |
|---|---|---|
| 1 | %x [-1,0] | 政府 |
| 2 | %x [0,0] | 顺利 |
| 3 | %x [1,0] | 恢复 |
| 4 | %x [-1,0] /%x [-1,1] | 政府 n |
| 5 | %x [0,0] /%x [0,1] | 顺利 ad |
| 6 | %x [1,0] /%x [1,1] | 恢复 v |

其中前三个模板用于第一层条件随机场的训练，后三个模板用于第二层条件随机场的训练。对于“中国/ns 政府/n 顺利/ad 恢复/v ”这个词语序列，在当前词为“政府”时，各特征模板扩展出的上下文特征如表 2 第三列所示。另外要说明的是，在第一层和第二层的条件随机场训练时均用到了词性转移特征模板。

## 2.2 性能评估

在评估汉语词性标注性能时，采用 Bakeoff 中常用的评测指标：标注精度（Accuracy）。标注精度表示已正确标注词性的词语在全部词语的标注词性中所占的比值。计算公式如下：

$$\text{Accuracy} = \frac{\text{正确标注词性的词语数}}{\text{所有待标注词性的词语数}} \quad (4)$$

## 2.3 实验结果及其分析

设计好特征模板后，在 Bakeoff2007 提供的训练语料上进行了训练。由于没有标准测试语料，实验过程中分别在两种语料的后面截取一部分作为测试语料，剩余的部分作为训练语料。在两种语料上进行训练时得到的相关数据如表 3 所示。

**表 3 NCC 和 CTB 语料上的训练过程记录数据**

Table 3 Record of Training Process on NCC and CTB

| 训练层数 | NCC 语料训练数据 | | | | CTB 语料训练数据 | | | |
|---|---|---|---|---|---|---|---|---|
| | 训练数据大小（MB） | 特征数 | 训练时间（s） | 模型大小（MB） | 训练数据大小（MB） | 特征数 | 训练时间（s） | 模型大小（MB） |
| 1 | 3.20 | 2597952 | 15063.14 | 13.5 | 3.89 | 3570780 | 14883.95 | 18.4 |
| 2 | 4.15 | 14980926 | 114314.31 | 66.4 | 5.43 | 9078172 | 51662.56 | 42.9 |

表中 3 的训练层数指的是第一层条件随机场训练还是第二层条件随机场训练，两层训练过程中用到的原始训练数据是一样的，只是在第二层训练时在训练数据中增加了第三列的词性细分结果。分析表 3 中的数据可以看出，训练数据的大小和训练过程中产生的特征数不成正比，比如在第二层训练中，NCC 的训练语料小于 CTB 的训练语料，但训练时产生的特征数却远远大于 CTB 语料产生的特征数，而两种语料在训练时采用的特征模板又是一样的，这主要就是因为 NCC 语料中的词性标记的种类数要大于 CTB 语料中词性标记的种类个数。但采用双层条件随机场进行训练时，两次训练过程均能够顺利进行。在两次训练完成后，对两次训练得到的模型均进行了测试，其测试结果如表 4 所示。

**表 4 词性标注结果**

Table 4 Results of part - of - speech tagging

| 模型层次数 | NCC 语料 | | CTB 语料 | |
|---|---|---|---|---|
| | 测试语料（KB） | Accuracy（%） | 测试语料（KB） | Accuracy（%） |
| 1 | 470 | 93.44 | 665 | 91.36 |
| 2 | 478 | 90.91 | 764 | 90.19 |

由于在第一阶段中的标注错误在第二阶段中有可能被放大，比如在第一层训练中如果将一个名词错标为了动词，则在第二阶段进行细分时一定也会产生错误。为了评测第一阶段模型对第二阶段模型的影响，对两个阶段的模型均进行了测试并进行了评估，从表 4 中的数据也可以看出，第一阶段的标注结果确实影响到了第二阶段的标注，但是总体来说影响并不是太大。

# 3 结论及进一步研究展望

汉语词性标注是中文信息处理领域中一项重要的基础研究课题，本文深入分析了条件随机场对汉语词性标注建模时上下文特征的表示方法和上下文特征产生的内在机理，在此基础上提出了基于双层条件随机场的词性标注方法，并采用 CRF + + 工具包在 Bakeoff2007 提供的 NCC 和 CTB 两种语料上进行了训练和测试，实验证明该方法是可行的。但从表 3 中的实验数据可以看出，在第二阶段的建模过程中，产生的特征数还是比较多的，这就导致训练的时间比较长，因此能否进一步利用层叠条件随机场进一步减少每层建模中的特征数，进而优化汉语词性标注的建模过程将是下一步的研究重点。

**[参考文献]**

[1] 姜维，王晓龙，关毅，等. 基于多知识源的中文词法分

析系统 [J]. 计算机学报，2007，30(1)：137 - 145.
[2] Brill Eric. Transformation - based error - driven parsing [A] / /Proceedingsof the third International Workshop on Parsing Technologies [C]. Tilburg, Netherlands, 1993.
[3] 赵岩，王晓龙，刘秉权，等. 融合聚类触发对特征的最大熵词性标注模型 [J]. 计算机研究与发展，2006，43(2)：268 - 274.
[4] 魏欧，吴健，孙玉芳. 基于统计的汉语词性标注方法的分析与改进 [J]. 软件学报，2000，11(4)：473 - 480.
[5] 刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析 [J]. 计算机研究与发展，2004，41(8)：1421 - 1429.
[6] 王丽杰，车万翔，刘挺. 基于 SVMTool 的中文词性标注 [J]. 中文信息学报，2009，23(7)：16 - 21.
[7] 姜维，关毅，王晓龙. 基于条件随机域的词性标注模型 [J]. 计算机工程与应用，2006，21：13 - 16.
[8] 洪铭材，张阔，唐杰，李涓子. 基于条件随机场(CRFs)的中文词性标注方法 [J]. 计算机科学，2006，33(10)：148 - 155.
[9] PEREIRA L J, MCCALLUM F A. Conditional random fields: probabilistic models for segmenting and labeling sequence data [A]. Proceedings of 18th Int Conf on Machine Learning. San Francisco [C]. USA: AAAI Press, 2001: 282 - 289.
[10] Nianwen Xue. Chinese Word Segmentation as Character Tagging [J]. Computational Linguistics and Chinese Language Processing, 2003, 8 (1): 29 - 48.
[11] 于江德，睢丹，樊孝忠. 基于字的词位标注汉语分词 [J]. 山东大学学报(工学版)，2010，40(5)：117 - 122.
[12] 冯冲，陈肇雄，黄河燕，等. 基于条件随机域的复杂最长名词短语识别 [J]. 小型微型计算机系统，2006，27(6)：1134 - 1139.
[13] 周俊生，戴新宇，尹存燕，等. 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，2006，34(5)：804 - 809.
[14] 于江德，周宏宇，余正涛，等. 基于单个词语特征模板的汉语词性标注 [J]. 山西大学学报(自然科学版)，2011，34(4)：513 - 517.

# The Quantitative Analysis of the Context Effective Range in Chinese Word Segmentation Based on Word Boundary Tagging

WANG Yi - fan[1], WANG Xi - jie[2]

(1. School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430000, China;
2. School of Computer and Information Engineering, Anyang Normal University, Anyang 455002, China)

**Abstract**: Chinese part - of - speech tagging often has the problem of too many well defined lexical catalogs. To improve this problem, the paper proposes a Chinese part - of - speech tagging method based on Dual - Layer conditional random fields. The approach divides the tagging procedure into two stages, each of which uses single - lyer conditional random fields to complete modeling. The first stage using context achieves coarse - grained part - of - speech tagging of each word. Taken the coarse - grained result as features, the second stage further produces sequences of fine - grained part - of - speech tags. Closed evaluations are performed on NCC and CTB corpus from the Bakeoff - 2007, and comparative experiments are performed on different feature templates. Experimental results show that this approach can obtain better pos tagging set.

**Key words**: Chinese part - of - speech tagging; Dual - layer conditional random fields; Context; Feature templates; Coarse - grained part - of - speech tagging

[责任编辑：江雪]