

# InsunPOS: 基于条件随机域的词性标注系统\*

杨小锐 刘秉权 孙承杰 林磊

(哈尔滨工业大学计算机学院智能技术与自然语言处理实验室)

{Xryang, liubq, cjsun, lin}@insun.hit.edu.cn

**摘要:** 本文描述了一个以条件随机域模型为主分类器、同时融合最大熵模型的词性标注系统。该系统参加了2009年的中文信息学会句法分析评测(CIPS-ParsEval-2009),在词性标注封闭测试中,正确率为93.3%,排名第二;在开放测试中,正确率为93.4%,排名第一。

**关键词:** 词性标注; 条件随机域; 最大熵模型; 隐马尔科夫模型

## InsunPOS: a CRF-based POS Tagging System

Xiaorui Yang, Bingquan Liu, Chengjie Sun, Lei Lin

(ITNLP lab, School of Computer, Harbin Institute of Technology)

{Xryang, liubq, cjsun, lin}@insun.hit.edu.cn

**Abstract:** This paper describes a pos tagging system InsunPOS. In InsunPOS, Conditional Random Fields (CRFs) are employed as the primary model; Maximum Entropy (ME) is also adopted to improve the performance of the systems. The system participated in the POS track of CIPS-ParsEval-2009 open evaluation. In the close test, InsunPOS achieved an accuracy of 93.3% and got the second place; and in the open test, InsunPOS achieved an accuracy of 93.4% and ranked 1<sup>st</sup>.

**Keywords:** Part-of-Speech Tagging; Conditional Random Fields; Maximum Entropy model; Hidden Markov Model

## 1 引言

词性标注是为句子中每一个词赋予正确的词性标记,它是许多深层自然语言处理任务必需的前序步骤,如组块分析、句法分析等。作为这些应用的预处理,词性标注中出现的错误将级联传入到后续处理中,直接影响到机器翻译、信息抽取以及问答系统等应用的性能。

常用的词性标注方法主要有以下几种:基于规则的方法,如Transformation Based Learner (TBL)方法[1];统计决策树(SDT)法[2];以及基于统计的方法,如隐马尔科夫(HMM)模型[3],最大熵(ME)模型[4, 5]以及支持向量机(SVM) [6]模型。基于规则的方法适应性较差,并且非统计模型的本质使得它不能给出每种可能分类结果的概率值,因此很难被用作一个更大概率模型的组件部分;SDT 虽然可以加入丰富的特征,但是在处理基于词的自然语言问题

---

\*国家自然科学基金面上资助项目(60673019, 60673037); 国家 863 计划资助项目(2007AA01Z172)

时, 需要数据划分, 因而存在严重的数据稀疏问题, 在应用时需要复杂的平滑技术; HMM是产生式模型, 存在两个主要缺点: 一是需假设特征之间彼此独立; 二是使用联合概率来模拟条件概率模型, 而实际自然语言处理问题往往是条件概率问题, 这使得HMM难于加入新的特征[7]。相比之下, 条件概率模型不要求特征独立性假设, 并且允许增加各种颗粒度的特征。条件随机域(CRF)模型是当前被广泛采用的用来解决序列化标注问题的一种条件概率模型。文献[8]表明采用CRF模型来进行词性标注, 其结果要优于HMM模型。

本文所描述的词性标注系统也采用了CRF模型, 引入的特征包括上下文特征和词的前、后缀等。为了进一步提高词性标注的准确率, 在close test中, 我们采用了CRF模型的标注结果。在open test中, 我们还进一步融合了最大熵模型的标注结果。该系统在封闭测试中, 准确率为93.3%, 排名第二; 在开放测试中, 准确率为93.4%, 排名第一。

## 2 系统描述

### 2.1 语料准备

我们对评测主办方提供的训练语料进行了划分。本次评测发放的语料分为 NEWS、BAIKE、HYL 三种类型, 每种类型的语料被随机分为五等份, 其中四等份作为训练集, 剩下的一份作为开发集。

在开发过程中, 我们把词分为单类词、兼类词、未登录词等三种, 单类词和兼类词的并集称为词表词。单类词是指词形在词表中出现、词性唯一的词; 兼类词是指词形在词表中出现、但具有多个词性的词; 未登录词包括以下两种情况: 一是当前词的词形未在词表中出现, 二是当前词的词形虽然在词表中出现, 但是其相应的词性却没有出现。

### 2.2 实验基线

我们用本实验室已有的基于 HMM 模型的词性标注系统作为 baseline 来估计本次评测语料的难度。采用 2.1 节所描述的训练集进行训练, 开发集进行测试, HMM 模型的总体准确率为 89%, 其中未登录词准确率为 43%, 兼类词准确率为 84%。这说明本次词性标注的语料是具有一定难度的。

### 2.3 条件随机域词性标注器

条件随机域是序列标注模型。设随机变量 $X$ 和 $Y$ 分别代表输入的句子与对应的词性序列, CRF模型通过局部特征向量 $f$ 与相应的权重向量 $\lambda$ 来表示。在CRF模型中,  $f$ 被分为状态特征 $s(y, x, i)$ 与转移特征 $t(y, y', x, i)$ , 其中 $y$ 与 $y'$ 是可能的词性标记,  $x$ 是当前的输入句子,  $i$ 是当前词的位置[9]。其形式化表述如下:

$$s(y, x, i) = s(y_i, x, i) \quad (1)$$

$$t(y, x, i) = \begin{cases} t(y_{i-1}, y_i, x, i) & i > 1 \\ 0 & i = 1 \end{cases} \quad (2)$$

根据公式①与②的局部特征，输入句子x与词性序列y的全局特征为：

$$F(y, x) = \sum_i f(y, x, i) \quad (3)$$

此时在(X, Y)上，CRF的条件概率分布为：

$$P_\lambda(Y | X) = \frac{\exp(\lambda \cdot F(Y, X))}{Z_\lambda(X)} \quad (4)$$

其中  $Z_\lambda(X) = \sum_y \exp(\lambda \cdot F(y, x))$  为归一化因子。

对于输入的句子x，最佳词性标注序列 $y^*$ 满足如下公式：

$$y^* = \arg \max_y p_\lambda(y | x) \quad (5)$$

## 2.4 特征

我们利用CRF++工具包进行模型的训练和测试。CRF词性标注器对一句话的标注流程如图1所示。

为了提高词性标注的性能，我们采用了后引入词典的方式，对CRFs模型探测出的词性进行了如下的调整：①若当前词是单类词（相对于词典来说），当且仅当CRFs模型探测出的词性的得分高于0.9时，采用CRFs模型探测出的词性作为最终的探测词性；否则以该词在词典中的词性作为最终的探测词性。②若当前词是兼类词，当CRFs所探测出的词性为该词词表词性之一时，采用CRFs探测出的词性作为最终的探测词性；当CRFs所探测出的词性不为词表词性且该词性的得分超过0.9时，此时采用CRFs探测出的词性作为最终的探测词性；当CRFs所给出的词性不为词表词性且该词性的得分低于0.9时，认为词表词性中得分排名第一的词性为最终的探测词性。③若当前词是未登录词，首先利用规则进行识别，如果将其识别为某种词性，那么将该词性作为最终的探测词性；否则，采用CRFs探测出的词性作为最终的探测词性。

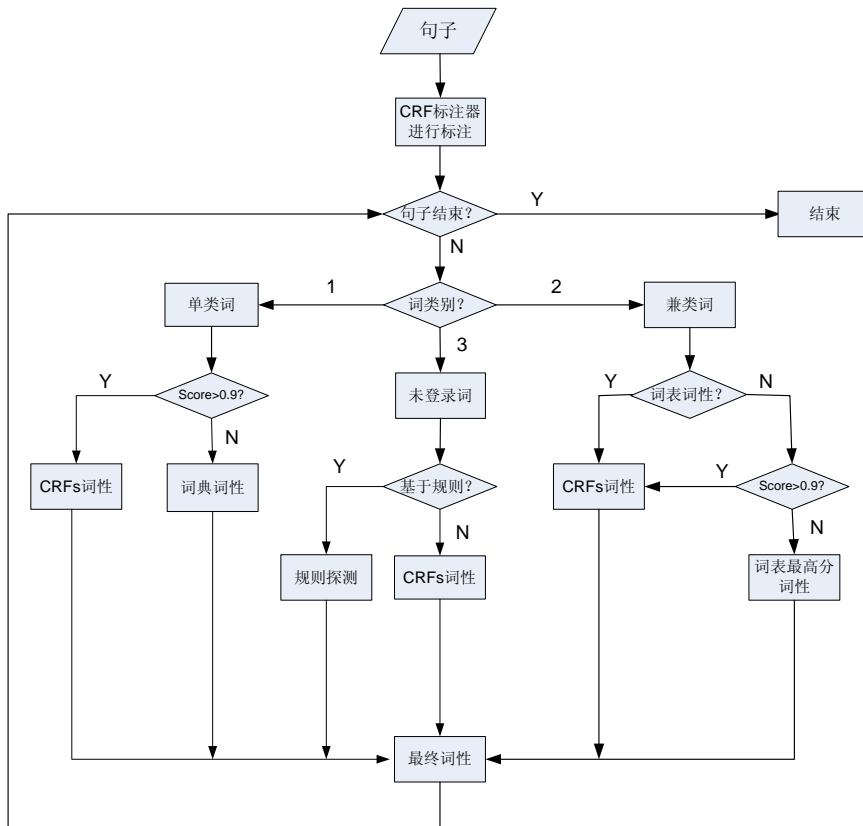


图 1 CRF词性标注器对一句语料的标注流程

我们先测试了表1所示的特征的效果。仍然采用2.1节所描述的训练集进行训练，开发集进行测试，条件随机域模型的准确率为92.2%，相比于HMM模型提高了3.2个百分点。这表明，CRFs相比HMM来讲，确实更适合处理词性标注这种序列化标注问题。根据对测试结果的统计，我们发现此时未登录词的标注准确率为48.7%（开发集中未登录词大约占6%）。因此提高未登录词的标注准确率将有助于提高整体的词性标注的准确率。以下两种方法被用来提高未登录词的标注准确率：一种是基于规则，另一种是在CRF模型中引入更多的特征（注：我们将利用表1特征模板训练的模型记作CRFs\_1。）

表 1 特征模板一

| 特征模板  |
|---|
| $w_{i+2}, w_{i+1}, w_i, w_{i-1}, w_{i-2}, w_{i+1}w_i, w_iw_{i-1}$ |

注： $w_i$ 表示当前词、 $w_{i-1}$ 表示前一词、 $w_{i+1}$ 表示后一词，以此类推。

首先我们采用引入规则的方法。经过对错误标注数据的分析，我们发现在未登录词中存在着大量的人名、地名、机构名、数词等等。例如：人名“周文骏”、“邓中夏”；地名“阳

高县”、“青海省”；机构名“中国科学院”、“奥委会”；数词“2.5”、“30%”。而这些类词在词形上往往有着明显的特征。人名的姓氏分布相对集中，人名长度多为4或6个字节（采用GB2312编码）；地名的尾字多为“山”、“河”、“省”、“市”、“县”、“镇”、“村”等等；机构名的尾字多为“院”、“校”、“部”、“委”等等；我根据这些构词特征，利用正则表达式来加强对这部分词的词性探测。这里我们仅以解决人名识别为例：我们统计出训练语料中出现频率较高的姓氏，构成姓氏表，在识别过程中，我们采用的策略如下，当一个词的CRFs得分低于0.5时，如果此时，这个词的首字在姓氏表中，并且该词条的长度为4或6个字节，我们就将其识别为人名。经过规则的引入，对未登录词的识别准确率上升到54%，总体的准确率上升到92.8%。（注：我们将利用表1特征模板训练的、引入规则的模型记作CRFs\_2）

表2 特征模板二

| 特征模板  |
|---|
| $w_{i+2}, w_{i+1}, w_i, w_{i-1}, w_{i-2}, w_{i+1}w_i, w_iw_{i-1}, pre_i, pre_iw_i, suf_i, w_iw_{i-1}$ |

通过对错误标注数据的进一步观察分析，我们发现在训练语料中存在“红灯”这个词，而在测试语料中的“走马灯”却标注错误。因此，我们设想可以通过引入词的首字、尾字等词缀特征来提高标注的准确率。采用表2的特征模板，CRF模型的标注准确率提高到93.39%。未登录词的标注准确率提高到58.8%。这表明利用词的前缀和后缀信息确实能够提高未登录词和兼类词的标注准确率。我们利用pre表示词的前缀，这里取该词的第一个字，suf表示词的后缀，这里取该词的最后一个字。例如：若当前词为“走马灯”，那么pre为“走”，suf为“灯”。（注：我们将利用表2特征模板训练的、引入规则的模型记作CRFs\_3）

## 2.5 引入外部资源

在对模型CRFs\_2标注的结果进行分析，虽然通过引入了规则大幅度的提高了人名、地名、机构名的标注准确率，但是结果仍然不理想，因此我们设想通过利用基于最大熵的名实体识别程序来对这三类词进行标注。我们的策略是首先利用CRFs\_2进行标注，如果该词条不在词典之中，并且名实体识别程序将该词识别为人名、地名、机构名、或其他专名之一时，我们采用名实体识别程序给出的标注结果作为最终的词性。实验表明，总体准确率由92.8%提升到93.3%，增幅为0.5个百分点。（注：我们将该模型记为O-CRFs\_2）

## 3 实验结果及分析

我们采用四份语料作为训练集、一份语料作为开发集。利用CRF++工具包进行训练，在P4、3.0GHz、1G内存环境下，模型所需的训练时间为75小时左右。各个模型的标注结果如

表 4 所示。

表 4 各个模型的标注结果

| 模型       | 准确率    |
|----------|--------|
| HMM      | 89%    |
| CRFs_1   | 92.2%  |
| CRFs_2   | 92.8%  |
| 0_CRFs_2 | 93.3%  |
| CRFs_3   | 93.39% |

我们利用全部数据进行训练。得到模型 CRFs\_3，在此模型基础上引入外部资源，该模型记为 0\_CRFs\_3。根据主办方发布的结果，利用模型 CRFs\_3 标注的结果准确率达到 93.3%，在封闭测试中排名第二；利用模型 0\_CRFs\_3 标注的结果准确率达到 93.4%，在开放测试中排名第一。实际提交系统的结果如表 5 所示。

表 5 评测结果

| 测试类型 | 模型       | 准确率   | 排名 |
|------|----------|-------|----|
| 封闭测试 | CRFs_3   | 93.3% | 2  |
| 开放测试 | 0_CRFs_3 | 93.4% | 1  |

在开发过程中，通过引入基于最大熵的命名实体识别程序，会对整体有 0.5 个百分点的贡献，但是在实际评测时，只有 0.1 个百分点的贡献。分析可能的原因如下：开发过程中是在 CRFs\_2 模型的标注结果基础上引入命名实体识别程序，在实际评测时，是在 CRFs\_3 模型的标注结果基础上引入命名实体识别程序，根据前文的分析，CRFs\_3 模型比 CRFs\_2 模型的标注准确率高，因此也就相应的弱化了命名实体识别程序的贡献。

## 4 结论

本文描述了一个以 CRF 模型为主分类的汉语词性标注系统 InsunPOS。在构建 InsunPOS 系统的过程中，我们主要有以下几点发现：1) 词的前、后缀特征可以显著提高词性标注的准确率；2) 通过后引入词典的方式对 CRFs 模型的结果进行必要的重新标注可以提高词性标注的准确率；3) 通过引入外部资源，可以有针对性提高某类词的标注性能，进而提高整体标注的正确率。

## 参考文献

1. Eric Brill. Transformation-based Error-driven Learning and Natural Language

- Processing: A Case Study in Part-of-speech Tagging. *Computational Linguistics*, 1995, 21(4):543-565
2. David M Magerman. Statistical Decision-Tree Models for Parsing. In proceeding of the 33rd Annual Meeting of the ACL, 1995: 276-283
  3. 胡春静, 韩兆强. 基于隐马尔可夫模型(HMM)的词性标注的应用研究[J]. *计算机工程与应用*, 2002, 38(6): 62-64
  4. Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In proceedings of the Conference on Empirical Methods in Natural Language Processing, 1996: 133-141
  5. 屈刚, 陆汝占. 基于特征的汉语词性标注模型. *计算机研究与发展*, 2003, 40(4): 556-561
  6. Jesús Giménez and Lluís Márquez. SVMTool: A general POS tagger generator based on Support Vector Machines. In proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. 2004: 43-46
  7. 姜维. 统计中文词法分析及其强化学习机制的研究. 哈尔滨工业大学博士论文, 2007: 31-32
  8. J Lafferty, A McCallum, F Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning, 2001: 282-289
  9. 姜维, 关毅, 王晓龙. 基于条件随机域的词性标注模型. *计算机工程与应用*, 2006; 38(21): 13-16