中文分词和词性标注联合模型综述

赵芳芳，蒋志鹏，关 毅

（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： 中文分词和词性标注任务作为中文自然语言处理的初始步骤，已经得到广泛的研究。由于中文句子缺乏词边界，所以中文词性标注往往采用管道模式完成：首先对句子进行分词，然后使用分词阶段的结果进行词性标注。然而管道模式中，分词阶段的错误会传递到词性标注阶段，从而降低词性标注效果。近些年来，中文词性标注方面的研究集中在联合模型。联合模型同时完成句子的分词和词性标注任务，不但可以改善错误传递的问题，并且可以通过使用词性标注信息提高分词精度。联合模型分为基于字模型、基于词模型及混合模型。本文对联合模型的分类、训练算法及训练过程中的问题进行详细的阐述和讨论。

关键词： 中文分词；中文词性标注；联合模型

中图分类号： TP391 **文献标识码：** A **文章编号：** 2095－2163(2014)03－0077－04

The Review on the Joint Model of Chinese Word Segmentation and Part－of－speech Tagging

ZHAO Fangfang, JIANG Zhipeng, GUAN Yi

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Chinese word segmentation and part－of－speech (POS) tagging task as an initial step for Chinese natural language processing, has been widely studied. Due to the lack of Chinese sentences word boundary, the Chinese POS tagging task is often completed with the pipeline approach: firstly, perform Chinese word segmentation, and then use the results of the prior stage to tag the Chinese sentence. However, in the pipeline approach, word segmentation phase errors will be passed to the POS tagging stage, thereby reducing the accuracy of POS tagging. In recent years, the research on Chinese POS tagging focused on the joint model. The joint model perform both word segmentation and POS tagging in a combined single step simultaneously, through which the error propagation can be avoided and the accuracy of word segmentation can be improved by utilizing POS information. There are character－based methods, word－based methods, and hybrid methods. In this paper, the three kinds of joint model, the training algorithm and the problems through the processing will be introduced in detail.

Key words: Chinese Word Segmentation; Chinese Part－of－speech Tagging; Joint Model

0 引 言

由于中文词与词之间没有自然界定，分词即成为中文自然语言处理的必要步骤。已经陆续研发了一些机器学习方法可用以解决分词和词性标注任务，譬如，隐马尔可夫模型[1]（Hidden Markov Mode，HMM）、最大熵模型[2]（Maximum Entropy Model，ME）及条件随机场[3]（Conditional Random Fields，CRFs）。具体地，CRFs在这两个任务中的表现堪称最佳。此外，感知器算法[4]（Perceptron Algorithm）也是一种广泛使用的判别式方法，该方法在大量训练语料的情况下，与CRFs的表现相当，而且在训练时间上更要低于CRFs方法。

为了建立词性标注器，通常有两种策略：

(1) 管道模型。先分词，再进行词性标注；

(2) 联合模型。分词和词性标注任务同时进行。

近些年来，联合模型方面的研究正大量涌现[5~14]。相应的研究已经表明，使用联合模型可以有效地降低错误传递，并且有助于使用词性标注信息实现分词，而为了提高分词和词性标注任务的准确率[5~6]，其代价将会是更大的搜索空间，更长的训练及解码时间。联合模型分为三种，分别是：基于字的联合模型、基于词的联合模型及混合模型。多类模型的研究实验表明，基于字的模型表现要优于基于词的联合模型[5]。目前，联合模型在中文开发领域已经取得了良好的精度，尤其是，PCTB语料中，分词 F 值可达98.44%[14]，词性标注 F 值可达94.17%[13]。

本文介绍了三种联合模型：基于字模型、基于词模型及混合模型，并且对联合模型中常用的训练和解码算法以及经常遇到的问题进行了总结。此后，介绍了不同于管道模型的评价方法，更进一步地则比较了管道模型及联合模型的表现。

1 中文分词和词性标注联合模型

1.1 联合模型概述

根据噪声信道模型的原理，可以将分词和词性标注的联

收稿日期：2014－05－09
基金项目：国家自然科学基金(60975077)。
作者简介：赵芳芳(1990－)，女，河南许昌人，硕士研究生，主要研究方向：自然语言处理；
蒋志鹏(1985－)，男，黑龙江七台河人，博士研究生，主要研究方向：自然语言处理；
关 毅(1970－)，男，黑龙江宁安人，博士，教授，博士生导师，主要研究方向：用户健康信息学、网络挖掘、自然语言处理等。

合模型对应的问题描述为:一个已经被标注了词性的词序列 $<W,T>=(<w_1,t_1>,<w_2,t_2>,\cdots,<w_n,t_n>)$(其中 $<w_i,t_i>(i=1,2,\cdots,n)$,表示具有词性 t_i 的词 w_i,下标 n 代表词串中词的个数),即词-词性标注对序列 $<W,T>$,经过有噪声的信道,将词边界和词性信息丢失,而在出口端输出为字序列 $C=(c_1,c_2,\cdots,c_{|C|})$(其中,$c_i$ 表示字串 C 中第 i 个字,$|C|$ 表示字串 C 的长度,并且 $\sum_{i=1}^{n} w_i=|C|$),通过找到与 C 相关联的 $<W,T>$,经比较得出具有最大概率的标注结果 $\hat{R}$,其计算过程如公式(1)所示,而分词和词性标注联合模型即如图1所示。

$$\hat{R}=<W,T>=\arg\max_{<W,T>}P(<W,T>|C) \quad (1)$$

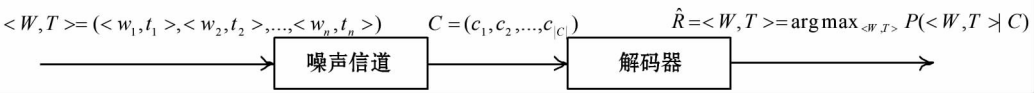

图1 分词和词性标注联合模型

Fig. 1 The Joint Model of word segmentation and part-of-speech tagging

1.2 联合模型分类

联合模型主要分为使用单一模型的联合方法以及混合模型,其中的前者又可分为基于字的联合模型和基于词的联合模型两类。在联合模型中通常会利用文献[4]中提出的感知器模型,并将其作为训练及解码算法,而在混合模型中则还会采用各种不同的方式,以此引入更多训练特征。

1.2.1 使用单一模型的联合方法

该方法包括两种不同的处理方式,也就是基于字的方法与基于词的方法,这两者的区别主要在于标注过程中对句子的基本处理单位不同,一者为字,另一者为词。

具体来说,基于字的联合模型指的是,对一个输入字串,其基本处理单位为字,因此就需要对每个字都标注位置标签及词性标签,再通过解码算法,运算得到最有可能的标注序列。

基于词的联合模型指的是,对于一个输入字串,其中的每个字,都要判断该字所属词的边界,确定得到一个完整的词后,再标注该词的词性标签。因此可知,这类模型中需要不停地判断当前字是否独立成词,或者在多字词中的位置,这就决定了需要用到词典信息。

在训练和解码算法中,两种模型实现过程的主要差异就在于训练模板以及进行词性标注时是以字,还是以词为单位[5]。两者的训练和解码算法一致。下面将集中介绍训练过程,而解码和训练算法则是相同的。训练阶段的流程图如图2所示。

训练阶段中,训练模板的选择对联合模型的表现具有很大影响,本文列举文献[5]中的基于字模型和基于词模型的部分关键特征模板,详细介绍如表1和表2所示。

表1 基于字模型的特征模板

Tab. 1 The feature templates of the character-based model

基于字模型的特征模板	
(a)	$C_n(n=-2,-1,0,1,2)$
(b)	$C_nC_{n+1}(n=-2,-1,0,1)$
(c)	$C_{-1}C_{-2}$
(d)	W_0C_0
(e)	$POS(C_{-1w_0})$
(f)	$POS(C_{-2w_0})\ POS(C_{-1w_0})$

表2 基于词模型的特征模板

Tab. 2 The feature templates of the word-based model

基于词模型的特征模板	
(a)	$W_n(n=-2,-1,0,1,2)$
(b)	$W_nW_{n+1}(n=-2,-1,0,1)$
(c)	$W_{-1}W_{-2}$
(d)	$POS(W_1)$
(e)	$POS(W_{-2})\ POS(W_{-1})$

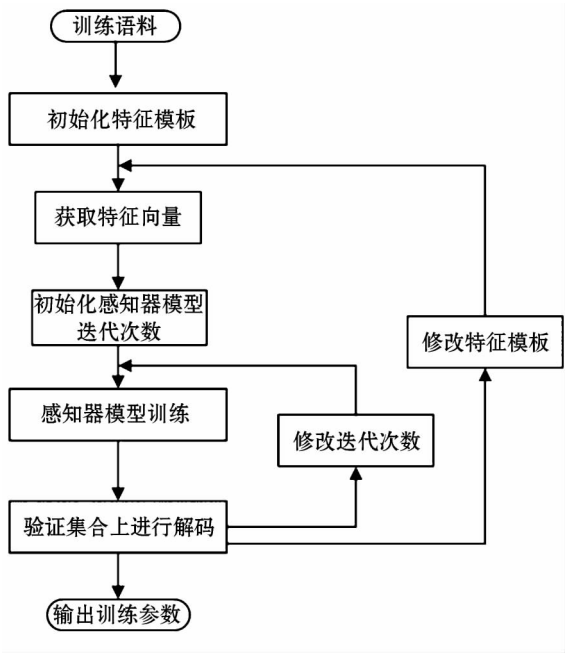

图2 训练阶段流程图

Fig. 2 The data flow of the training stage

本文使用了感知器算法训练模型,作为线性模型,其可处理的是实数空间的变量。因此使用该模型进行词性标注前,需要将特征离散为数值。与文献[4]类似的是,本文中 ϕ 为 d 维向量,可形式化表示为 $\phi: H\times T\rightarrow R^d$($R^d$ 表示 d 维实空间),向量的每一维分量 ϕ_i(其中,$1\leqslant i\leqslant d$)即为指示函数,可将该维特征离散化为实数。

在此使用 Φ 来表示序列($C_{[1:n]}$,$t_{[1:n]}$)对应的 d 维向量,可将 Φ 称为全局向量,ϕ 为局部向量,两者的关系如下:

$$\Phi_d(C_{[1:n]}, t_{[1:n]}) = \sum_{i=1}^{n} \phi_d(h_i, t_i) \quad (2)$$

其中，Φ_d 与 ϕ_d 分别表示相应函数的第 d 维分量。当 ϕ_d 为指示函数时，Φ_d 即表示一个句子出现第 d 维特征的次数。

训练算法如图 3 所示，算法输入为 n 行训练语料及迭代次数 N。输出为参数向量 α。

感知器算法

输入: 训练样本，总行数为 n; 迭代次数 N

初始化: 参数向量 $\alpha = 0$

训练过程:

for $t = 1..N$, $i = 1..n$

●使用本文的 Viterbi 算法找到当前参数下满足下面条件的标注序列

$$y_{[1:|C_i|]} = \text{argmax}_{y \in \Gamma^{C_i}} \sum_d \alpha_d \Phi_d$$

●如果 $y_{[1:|C_i|]} \neq t^i_{[1:|C_i|]}$ 则更新参数向量的每一维

$$\alpha_d = \alpha_d + \Phi_{d\_gold} - \Phi_d$$

输出: 参数向量 α

图 3 联合模型训练算法

Fig. 3 The training algorithm for joint model

图 3 中，Γ^{C_i} 表示输入字串 C_i 后所有可能的输出结果，即 $y, \hat{y}, y^* \in \Gamma^{C_i}$。其中的 $\hat{y}$ 与 y^* 分别表示模型得到的最可能标注结果与标准标注结果。

1.2.2　混合模型

混合模型中采用了多于一个的模型，也就是通过使用更多的特征来表示上下文信息。混合模型中常常将基于字的模型和基于词的模型结合起来加以使用，并且还有可能会结合 n - gram 语言模型及最大熵等模型。混合模型的一个难点是如何将多个模型有效地结合在一起。

文献 [8] 提出一种基于词网格的重排方法，将基于字的联合模型与基于词的联合模型串联起来，具体方法为: 首先在训练语料上训练基于字的联合模型，接着，使用词网格方法标注训练语料，最后，训练基于词的模型，但该模型仅考虑词网格中的词; 文献 [12] 又提出堆叠式学习方法(Stacked Learning Method) ，是将基于词的联合模型与基于词的联合模型并联起来进行处理。处理方法是，在同一个训练集上，使用相同的交叉验证方法训练基于字的模型与基于词的模型，同时对得到的结果赋予不同的权重。文献 [14] 则结合使用了词网格方法及堆叠式学习方法。

然而，使用的特征越多，训练和解码所需要的时间越长。由于联合模型是生成模型，因此在训练过程中将产生大量的无意义的词，这就给模型的训练和解码带来了困难。针对这一问题，文献 [14] 中提出了一种减少无意义词的方法，实验结果表明，减少的无意义词已经可以达到 62.9%。文中更进一步地提出了一种可靠的成词标准，将其表述为: (1) 添加通用词典; (2) 使用维基百科中记录的命名实体; (3) 添加大规模未标注的源语料。

2　实验与分析

2.1　评价方法

联合模型与管道模型中的评价方法是各不相同的，下面介绍的是文献 [5 - 14] 中的通用评价方法。该方法可做如下描述:

当且仅当该词边界同时出现在 $\hat{y}$ 和 y^* 中，表示分词正确; 当且仅当该词边界及该词的词性标签同时出现在 $\hat{y}$ 和 y^* 中，表示词性标注正确。

联合模型中分词和词性标注 F_1 的定义如下:

$$F_1 = \frac{2\sum_C |\hat{y} \cap y^*|}{\sum_C |\hat{y}| + \sum_C |y^*|} \quad (3)$$

其中，$|\hat{y} \cap y^*|$ 表示句子 C 中正确的词边界(评价分词阶段质量) 或正确的词标注的数量(评价词性标注阶段质量) 。

2.2　实验结果及分析

本节列举部分联合模型及管道模型在相同 PCTB 语料上的实验结果，具体如表 3 所示。其中，Baseline 为管道模型，与文献 [9] 中的 Baseline 相同。Baseline 中使用的训练模板及训练算法与文献 [9] 中的联合模型(Zhang and Clark，2008) 一致，通过对比这两个模型即可得到联合模型的精度高于管道模型的结论。

表 3　不同模型对比

Tab. 3 The comparison of different models

词性标注器	分词(F_1)	词性标注(F_1)
(Zhang and Sun，2013)	0.981 1	0.938 3
(Zhang and Clark，2010)	0.977 8	0.936 7
(Jiang et al. ，2008b)	0.977 4	0.933 7
(Zhang and Clark，2008)	0.959 0	0.913 4
Baseline	0.952 0	0.903 3

3　结束语

本文主要阐述了联合模型的分类及训练过程，并且通过实验结果展示了联合模型与管道模型相比后而在分词和词性标注精度上获得的提升。然而，精度的提升却需要空间和时间上的代价。联合模型训练中所需的存储空间及时间都比管道模型有较大增加，并且会随着训练语料的增多更呈指数增长。解码阶段所需时间也比管道模型长。因此，当需要完成分词和词性标注任务时，就要考虑具体项目的空间和时间需求，并在此基础上再选择管道模型，或是选用联合模型。

更多实验结果显示，基于字的模型比基于词模型的表现要好，而混合模型又比基于字及基于词的效果都更好。在多模型混合的方法上，文献中也提供了不同的方法，目的是为了更好地结合多个特征。针对生成模型在学习过程中产生大量无意义词的问题，文献中也提供了适用的方法。

文中的联合模型可以用来解决跨领域问题。跨领域问题中的模型精度是影响精度的一个重要因素，因此使用联合模型即为提高跨领域标注的精度提供了一个新的研究思路。

参考文献:

[1] RABINER L R. A tutorial on hidden markov models and selected applications in speech recognition [C] //Proceedings of IEEE, 1989: 257 - 286.

[2] RATNAPARKHI, ADWAIT. A maximum entropy part – of – speech tagger [C] //Proceedings of the Empirical Methods in Natural Language Processing Conference,1996.

[3] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C] //Proceedings of the 18th ICML, 2001: 282 – 289.

[4] COLLINS M. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms [C] // Proceedings of EMNLP,2002: 1 – 8.

[5] NG H T, LOW J K. Chinese part – of – speech tagging: one – at – a – time or all – at – once? Word – based or character – based [C] // Proceedings of EMNLP, 2004.

[6] ZHANG Yue, CLARK S. Chinese segmentation with a Word – based perceptron algorithm [C] //Proceedings of ACL,2007, 8.

[7] JIANG Wenbin, HUANG Liang, LIU Qun,et al . A cascaded linear model for joint chinese word segmentation and part – of – speech tagging [C] //proceedings of the 46th Annual Meeting of the Association for Computational Linguistics,2008a.

[8] JIANG Wenbin, MI Haitao, LIU Qun. Word lattice reranking for chinese word segmentation and part – of – speech tagging [C] // proceedings of the 22nd International Conference on Computational Linguistics, 2008,1: 385 – 392.

[9] ZHANG Yue, CLARK S. Joint word segmentation and POS tagging using a single perceptron [C] //proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2008: 840 – 847.

[10] KRUENGKRAI C, UCHIMOTO K, KAZAMA J, et al. An error – driven Word – character hybrid model for joint Chinese word segmentation and POS tagging [C] // Proceedings of ACL – IJCNLP, 2009:513 – 521.

[11] ZHANG Yue, CLARK S. A fast decoder for joint word segmentation and POS – Tagging using a single discriminative model [J]. EMNLP – 10, 2010.

[12] SUN Weiwei. A stacked sub – word model for joint Chinese word segmentation and part – of – speech tagging [C] //Processings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,2011: 1385 – 1394.

[13] WANG Yiou, KAZAMA J, TSURUOKA Y, et al. Improving Chinese word segmention and pos tagging with semi – supervised methods using large auto – analyzed data [C] //Proceedings of 5th International Joint Conference on Natural Language Processing, 2011: 309 – 317.

[14] ZHANG Kaixu, SUN Maosong. Reduce meaningless Words for joint Chinese Word segmentation and part – of – speech tagging [J]. Computation and Language , arXiv: 1305.5918, 2013.

(上接第76页)

4 结束语

红外辐射电暖器智能控制系统的应用能有效避免使用者因突发停电离开,而未断开红外辐射电暖器电源,当其离开期间市电又恢复供应所可能导致的能源浪费,甚至红外辐射电暖器寿命受损乃至引发火灾事故等诸多状况的发生。基于 AT89C51 单片机为核心的红外辐射电暖器智能控制系统在保证性能稳定、工作可靠的前提下,硬件设计上尽量采用性价比较高的通用元器件,因而降低了制作成本。红外辐射电暖器智能控制系统与已有的红外辐射电暖器配套使用时,并不需要重新购置电暖器,这样在无形中节约了使用者的成本。红外辐射电暖器智能控制系统可以实现智能感应、手动定时、断电自锁、延时关机等功能,提高了电能的利用率和电暖器的安全效果,因此具有较高的实际应用推广价值[11]。

参考文献:

[1] 王靖. 电暖器卖场调查 [J]. 大众标准化,2010(12):26 – 28.

[2] 张志伟. 模数混合信号集成电路自动设计技术研究 [J]. 陕西理工学院学报(自然科学版),2013,29(4):25 – 29.

[3] 彭建,孙志江. 基于单片机控制的降雨量实时监测系统 [J]. 测控技术,2010,29(9):79 – 84.

[4] 张志伟. 基于 ZigBee 技术的日光温室环境参数监测系统设计 [J]. 安徽农业科学. 2011,39(26):16393 – 16394.

[5] 段现星,王晓侃. 基于单片机控制的车载酒精浓度检测仪设计 [J]. 测控技术,2013,32(8):1 – 3 + 7.

[6] 周立功,等编著. 增强型 80C51 单片机速成与实战 [M]. 北京: 北京航空航天大学出版社,2003.

[7] 沙占友,编著. 集成传感器应用 [M]. 北京: 中国电力出版社,2005.

[8] 张志伟. 一种远程矿井瓦斯浓度检测仪的设计 [J]. 煤矿安全,2011,42(2):78 – 80.

[9] 姚有峰,赵江东,郝诗平. 基于单片机技术的环境状态监测系统的设计 [J]. 测控技术,2012,31(1):105 – 108.

[10] 李想. MPEG – 4 视频解码器的设计与优化 [J]. 计算机时代,2013(1):6 – 8.

[11] 张志伟. 远红外线电暖器安全节能控制系统设计 [J]. 电子质量,2013(3):12 – 14.