ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 3,May 2017,pp. 636－642
DOI:10. 16337/j. 1004-9037. 2017. 03. 024

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax: ＋86-025-84892742

基于 CRFs 和歧义模型的越南语分词

熊明明[1]　李　英[1]　郭剑毅[1,2]　毛存礼[1,2]　余正涛[1,2]

(1. 昆明理工大学信息工程与自动化学院，昆明，650500；2. 昆明理工大学智能信息处理重点实验室，昆明，650500)

摘　要：通过对越南语词法特点的研究，把越南语的基本特征融入到条件随机场中(Condition random fields，CRFs)，提出了一种基于 CRFs 和歧义模型的越南语分词方法。通过机器标注、人工校对的方式获取了 25 981 条越南语分词语料作为 CRFs 的训练语料。越南语中交叉歧义广泛分布在句子中，为了克服交叉歧义的影响，通过词典的正向和逆向匹配算法从训练语料中抽取了 5 377 条歧义片段，并通过最大熵模型训练得到一个歧义模型，并融入到分词模型中。把训练语料均分为 10 份做交叉验证实验，分词准确率达到了 96. 55％。与已有越南语分词工具 VnTokenizer 比较，实验结果表明该方法提高了越南语分词的准确率、召回率和 F 值。

关键词：条件随机场模型；越南语分词；词法；基本特征；最大熵；歧义模型

中图分类号：TP301　　**文献标志码**：A

Vietnamese Word Segmentation with Conditional Random Fields and Ambiguity Model

Xiong Mingming[1], Li Ying[1], Guo Jianyi[1,2], Mao Cunli[1,2], Yu Zhengtao[1,2]

(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China; 2. The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, 650500, China)

Abstract: The Vietnamese lexical features are discussed and essential characteristics of Vietnamese are integrated into condition random fields (CRFs) to propose a Vietnamese word segmentation method based on CRFs and ambiguity model. The segmentation corpus consisting of 25 981 Vietnamese is obtained as a training corpus of CRFs by computer marking and artificial proofreading. Vietnamese crossing ambiguity is widely distributed in the sentence. To eliminate the effects of crossing ambiguity, 5 377 ambiguity fragments are extracted from training corpus through dictionary of the forward and reverse matching algorithm. An ambiguity model is obtained by training the maximum entropy model. Then they are both incorparted into the segmentation model. The training corpus is divided into ten copies evenly for cross validation experiments. The segmentation accuracy reaches 96. 55％ in the experiment. Experimental results show that the method improves the segmentation accuracy rate, the recall rate and the F value of Vietnamese word obviously, compared with Vietnamese segmentation tool VnTokenizer.

Key words: condition random fields(CRFs); Vietnamese segmentation; morphology; essential characteristics; maximum entropy; ambiguity model

---

基金项目：国家自然科学基金(61262041，61472168，61562052)资助项目；云南省自然科学基金重点项目(2013FA030)资助项目。
收稿日期：2015-05-28；**修订日期**：2015-07-27

引　言

越南语分词是越南语信息处理的前提,是越南语词法、句法、语义以及各种上层应用的基础。目前在分词方面已经有很多研究成果,主要包括基于词典匹配的分词方法、基于统计的分词方法和基于理解的分词方法。基于词典匹配的方法主要按照一定的策略将待分析的字符串与一个"充分大的"机器词典中的词条进行匹配。若在词典中找到某个字符串,则匹配成功,把该词取出。如何国斌等用最大匹配法进行分析,达到了一定的效果[1]。基于统计的方法主要对语料库中词与词的组合进行统计,计算它们的互现信息。通过定义两个字符的互现信息,计算这两个字符相邻共同出现的概率。互现信息体现了字符之间结合关系的紧密程度。当紧密程度高于某一阈值时,便可认为此字符组合可能构成一个词,如张华平等使用基于层叠隐马模型的汉语词法分析方法,明显地提高了分词的准确率[2]。石民等把条件随机场用到先秦文字的分词中[3]。基于理解的方法是通过让计算机模拟人对句子的理解,达到识别词的效果。由于语言知识的笼统、复杂性,难以将各种语言信息组织成机器可直接读取的形式,因此目前基于理解的分词系统还处在试验阶段。在越南语分词的研究方面,国内外的研究都只是刚刚开始,直到现在还没有具体的共享资源可供学术研究使用,所有的语言资源都需要从头建起。越南语分词作为越南语自然语言处理的基础,要求收集越南语语料资源,并按照要求进行处理,这是越南语分词的前提条件。目前,VnTokenizer 是 2008 年由越南本国河内大学采用基于最大匹配和 N-Gram 模型开发的越南语分词工具[4]。本文在前人工作基础之上,结合越南语构词特征和语言特点,在条件随机场模型中融入了 N-Gram 模型、字符重复特征和字符类型特征,并加入歧义模型从而实现对越南语的分词。通过反复测试,并与 VnTokenizer 对比实验。结果表明该方法显著提高了分词效果。

1　越南语词法特点与特征提取

1.1　越南语的词法特点

越南语是一种有声调孤立语[5],与汉语拼音很相似,每一个音节也是由声母、韵母和声调组成。越南语的声母有 6 个,比如 a, à,ả, á,ā,ạ,分别为平声、锐声、玄声、问声、跌声和重声 6 个声调,又分为平、仄两类,其中前 2 个属于平,后 4 个属于仄。每一个音节几乎都有意义,越南语和汉语一样,缺乏形态变化[6]。除此之外,它的构成就是拉丁字母、表音文字和标点符号等。越南语的构词单位和汉语拼音是一样的,也是语素。越南语的构词可以分为 5 种,单音节词、复合词(并列复合和偏正复合)、重音叠韵词(完全重叠和部分重叠)、偶合词和派生词[7]。越南语的构词法见表 1。中空格英文的分词就是按照句子来分的,但是越南语分词不能按照空格来分,一个词可能有多个词素构成,比如:"Tôi là một sinh viên"的分词为:Tôi (我) /là (是)/ một (一个)/ sinh viên(学生)。这种由词素构词的词类似中文分词。

1.2　越南语歧义性

越南语有两种类型的歧义:组合歧义和交叉歧义。组合歧义中一些单独词素可以成词,这些词素合起来也成词,如:"Bàn là một công cụ học tập. (桌子是一个学习工具。)" 词素"Bàn"是"桌子"的意思,"là"是"是"的意思,而"Bàn là"又是"铁"的意思。这种歧义很难处理,但是越南语中这种歧义远远少于交叉歧义。交叉歧义就是当前词素与它的前一个词素和后一个词素都能成词。如:"Tốc độ truyền thông tin ngày càng cao. (传输信息速度越来越快)""truyền thông"和"thông tin"分别表示两个词

表 1　越南语构词规则

Tab. 1　Vietnamese word formation rules

构词类型	举例
单音节词	tôi 我 đẹp 漂亮
并列复合词	yêu thư,o,ng 爱护
偏正复合词	xe đạp 自行车
完全重叠	khăng khăng 偏偏
部分重叠	Dom đóm 萤火虫
偶合词	mit tinh(meeting)
派生词	ki-lô-gam→ ki lô(千克)

“媒体”和“信息”。这种歧义在越南语中经常发生，由于词典内容有限，很多未登录词难以消歧，是一种具有挑战性的问题。

1.3 越南语特征选取与特征模板的定制

对于统计模型基于条件随机场模型(Condition random fields，CRFs)，特征的选取对分词结果具有很大的影响，是关键环节。本文结合以上越南语的特点，在使用 CRFs 模型对越南语进行分词时，定义了两类特征。

(1)基本特征模板 1

在定制特征模版 1 时，选用了两类基本特征，这两类特征是 Tseng 提到的[8]：字符 N-gram 特征和字符重复信息特征，如表 2 所示。其中，W 代表越南语词素；W_0 代表当前词素，k 相对于当前词素所处的位置。比如：“Tôi không nói đu, ọ, c tiếng Việt. ”，如果 W_0 代表当前越南语词素“đu, ọ, c”；则 W_{-1} 表示“nói”；W_{-2} 表示“không ”；W_1 表示“tiếng ”，W_2 表示“Việt”。Repeat(W_0W_1)表示当前词素和下一个词素完全一样。

表 2 基本特征模板 1

Tab. 2 Basic feature template 1

特征	形式化描述
字符 N-gram 特征	$W_k(k=[-2,-1,0,1,2])$
	$W_kW_{k+1}(k=[-2,-1,0,1])$
字符重复信息特征	$W_kW_{k+2}(k=[-1,0])$
	$W_kW_{k+2}(k=[-1,0])$

(2)基本特征模板 2

针对越南语中的数字、字母和标点等容易出错的未登录词，本文根据语言特性将越南语词素定义为 10 大类别：Sin，Pre，Suf，Pun，Dig，Let，Spe，Tim，Dat，Oth[9]。本文所使用的词素类别的定义以及相关例子如表 3 所示。

表 3 基本特征模板 2

Tab. 3 Basic feature template 2

特征定义	特征表示	示例
单独成词	Sin	thi，ah，nó
词语的开始	Pre	học，c? m，xin
词语的结束	Suf	sinh，bạn，chào
标点符号	Pun	，. ！
数字	Dig	1，2，3
字母或字母组合	Let	A，a，CRF，BBC
特殊标识符	Spe	@ % …．．
时间	Tim	时分秒
日期	Dat	年月日
其他	Oth	Ⅰ Ⅱ 等

2 交叉歧义模型

最大熵模型是一个统计模型，被广泛地运用到自然语言处理和图像处理等领域[10,11]。它的特征选择灵活，建模时只需要集中精力选择特征，而不需要考虑如何使用，可以融入很丰富的信息。如果用 A，B，C 来表示交叉歧义片段，则考虑一下 4 类统计信息：(1)A 的独立成词概率是否大于 C；(2)A 与 B 成词概率是否大于 B 与 C；(3)A 作为词首的概率是否大于 C 作为词尾的概率；(4)B 作为词尾的概率是否大于 B 作为词首的概率。以上可以分别作为最大熵模型的模板，分别定义如下：If ($P(A)>P(C)$) Then 一阶模板 $T_1=1$ Else $T_1=0$；If ($P(AB>P(BC)$)) Then 二阶模板 $T_2=1$ Else $T_2=0$；If (P(A 首)$>P$(C 尾)) Then 三阶模板 $T_3=1$ Else $T_3=0$；If (P(B 尾)$>P$(B 首)) Then 四阶模板 $T_4=1$ Else $T_4=0$。为了构建交叉歧义模型，对 25 981 条越南语分词语料通过词典的正向匹配和逆向匹配方法获取了 5 377 条歧义片段。使用的词典含有 143 130 词条，最终形成的歧义模型的训练格式，如图 1 所示。

```
1  cơ ngơi ngốt  1  1  1  0
1  lộp độp một cái  1  0  1  0
1  máy xúc tiến  1  0  1  1
0  nuôi béo quay  0  1  1  1
1  não lòng đường  1  0  1  0
1  tứ khoái kéo  0  1  0  0
1  đen nhức nhối  1  1  1  1
0  xi nhan đê  0  1  1  0
0  mặt trăng trắng  0  1  1  0
```

图 1 歧义片段训练格式

Fig. 1 Training format of ambiguity fragments

第 1 列中的“1”表示切分方式为“A/BC”，“0”表示切分方式为“AB/C”。第 2 列是抽取的歧义片段，第 3～6 列分别表示一、二、三和四阶特征模板。

3 应用 CRFs 和歧义模型进行分词

3.1 CRFs 理论

鉴于条件随机场模型能够综合利用多层资源,同时在一定程度上能够避免歧义问题和数据标注偏执问题,本文采用 CRFs,它是一种机器学习模型,由 John Lafferty[12,13] 最早用于自然语言处理(Natural language process,NLP)领域的文本标注。近年来在分词、词性标注和命名实体识别等序列标注任务中也取得了很好的效果[14,15]。CRFs 是无向图的一种表现形式,在给定将要标注的观测序列的情况下,无向图模型可以被用来在标注序列上定义一个联合概率分布。假设 X、Y 分别表示需要标注的观察序列和它对应的标注序列的联合分布随机变量。对于给定的一个长度为 n 的序列,$X=x_1,x_2,\cdots,x_n$,则输出 $Y=y_1,y_2,\cdots,y_n$ 的概率可以定义为

$$P(Y/X)=\frac{1}{Z(x)}\exp\left\{\sum_{k=1}^{k}\boldsymbol{\lambda}_k f_k(y_t,y_{t-1},x_t)\right\} \tag{1}$$

式中:Z 为归一化常量,他使得所有的状态序列的概率和为 1。$Z(x)$的计算公式为

$$Z(x)=\sum_{y}\exp\left\{\sum_{k=1}^{k}\boldsymbol{\lambda}_k f_k(y_t,y_{t-1},x_t)\right\} \tag{2}$$

式中:$Z(x)$ 为归一化因子,$f_k(y_t,y_{t-1},x_t)$为对整个序列的 X 标记位于 i 和 $i-1$ 的特征函数,特征函数是一个二值函数,即布尔值,取值集合为{0,1}。$\boldsymbol{\lambda}_k$ 是每一个特征权重向量。在越南语分词系统中,条件随机场的训练就是通过训练语料,来学习最恰当的模型参数,来使得某种规则标准最大。在这里基于最大似然估计对条件随机场进行训练,使得条件概率的对数似然值最大。

3.2 分词系统

为了克服交叉歧义给分词带来的影响,本文在 CRFs 分词的过程中加入了交叉歧义模型,使其在分词结果准确率方面有所贡献。这里给出了分词的流程,如图 2 所示。越南语分词系统算法描述为:

输入:待分词句子 D ($D=\{S_1,S_2,S_3,\cdots,S_n\}$)

输出:分词结果

(1)首先使用词典的正向和逆向匹配算法找出带分词句子的歧义片段。

(2)如果没有歧义片段则执行步骤(4),如果有则把歧义片段放到数组中。

(3)循环数组,分别对数组中的歧义片段进行歧义切分,确定切分的结果是 A/BC 或者 AB/C。

(4)加载分词模型,对待分词句子进行切分。

(5)如果数组为空,则执行步骤(6),否则确定步骤(4)中的分词结果,用步骤(3)中的结果进行替换。

(6)输出最终的分词结果。

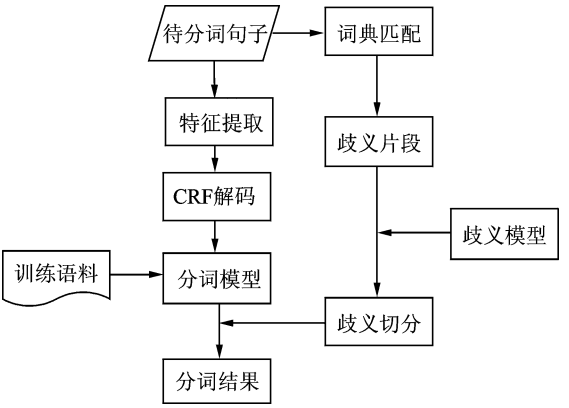

图 2 越南语分词流程图

Fig. 2 Flow chart of vietnamese segmentation

4 实验结果及分析

(1)实验语料的选择

本文采用的主要语料是通过在越南新闻网站爬取的越南语句子作为训练语料和测试语料,爬取的网页经过规则提取、去重、机器标注和人工校对等步骤形成文本语料库,其规模为 25 981 条句子,编码方式采用 UTF-8。

(2)实验测评标准

准确率和召回率是广泛用于信息检索和统计学分类领域的两个度量值,用来评价结果的质量。类似地,可以把这 3 种评价方法用到分词任务中,在越南语老师和留学生的帮助下,标注 25 981 条越南语句子,并做十倍交叉验证实验,训练得到的分词模型在交叉实验中进行测试。分词后的结果使用准确率 P,召回率 R 和 F 评价该分词系统。P=分词结果中切分正确的词数/分词结果中的总词数;R=分词结果中切分正确的词数/人工标注文本的总词数;$F=2*P*R/(P+R)$;其中准确率和召回率这两者在 0 和 1 之间,数值越接近 1,查准率或查全率就越高。F 即为准确率和召回率的调和平均值。

(3)分词系统性能测试实验

分词系统使用了"特征模板 1","特征模板 2"和"歧义模型"。对 10 份交叉验证的实验数据的测试结果如表 4 所示。表 5 给出了分词系统在测试集测试上的结果以及各个特征模板和歧义模型对结果的贡献。从表 4 可以看出:随着特征模板和歧义模型的逐个加入,分词的准确率、召回率和 F 逐步提高。其中,"模板 1+模板 2+歧义模型"模型的结果明显好于"模板 1+模板 2"模型的结果,即在增加了歧义模型的情况下,P 和 F 分别高出了 2.52%和 2.19%,可见歧义模型起到了较好的效果。而且由于歧义片段完全是从训练语料中提取的,最大熵模型可以很好地统计到歧义信息。

(4)与 VnTokenizer 的对比实验

为了进一步测试分词系统的实验效果,分别用分词系统与 VnTokenizer 对 1 000 条语料进行了测试实验,这 1 000 条测试语料的正确分词结果已经在越南语老师和同学的帮助下标注完成。实验结果对比如表 5 所示。从表 5 的对比实验发现:加入歧义模型的分词模型的准确率、召回率和 F 均有小幅度提升。利用训练语料生成分词模型,并对准确率、召回率和 F 进行比较分析可知,基于 CRFs 和歧义模型的越南语分词方法在分词方面取得了较好的效果。条件随机场对越南语构词特征具有较强的融合能力,随着特征集的增加,分词的效果会更好。

表 4 十倍交叉验证实验

Tab. 4 Tenfold cross-validation experiments %

特征模板	P	R	F
模板 1	90.52	90.03	90.27
模板 1+模板 2	94.03	93.89	93.96
模板 1+模板 2+歧义模型	96.55	95.76	96.15

表 5 分词实验结果对比

Tab. 5 Comparison of experimental results of Vietnamese segmentation %

系统	P	R	F
分词系统	96.86	95.95	96.40
VnTokenizer	95.52	96.03	95.77

(5)歧义词实验

针对歧义词的处理,首先考察分析了本文越南语分词系统词边界的消歧能力。由于歧义片断很难确定,所以目前只是简单地通过正向匹配分词和逆向匹配分词法对测试语料进行分词,然后通过双向比对来确定歧义片断。最后通过以下 3 种分词模型对测试语料进行分词,比较 3 种分词工具对歧义片段的切分结果,实验结果如表 6 所示。从表 6 可以看出,加入歧义模型的分词系统对歧义词的切分正确率显著提高,比没有加入歧义模型的分词系统的切分正确率提高了 10%,同时比 VnTokenizer 的切分正确率提高 5%。实验表明,歧义词的正确切分对句子分词的准确率有明显的提高作用。

(6)未登录词的处理实验

实验考察分析了 CRFs 对未登录词的识别能力。用分词模型对 1 000 条生语料进行测试。越南语分词若无法完成对未登录词的处理,或者其处理结果存在很大的误差,这样就会大大降低实际结果的准确性,因此导致统计出错误的词频统计信息、词频参量值。本文将未登录词分为 4 类:命名实体类、非越南词类(包括数字、年月日、大写字母和小写字母等)、外来衍生词和其他未登录词。命名实体类主要使用特征模板 1 很好地进行捕获。针对非越南语词类,主要通过字符类型特征进行区分,以达到识别的目

的。外来衍生词属于少数部分,几乎不能正确地切分,它不可能包含在训练语料中。对于其他未登录词,一般是指词素个数大于等于 4 的,CRFs 对其识别也有困难。在测试语料中未登录词数 302 个,切分正确 224 个,切分正确率为 74.17%。

表 6 歧义词实验结果

Tab. 6 Expermental result of Ambiguous words

模型	歧义词	切分正确	切分正确率/%
模板 1+模板 2	168	135	80.36
模板 1+模板 2+歧义模型	168	153	90.07
VnTokenizer	168	144	85.71

5 结束语

本文收集、整理了 20 万条越南语句子,通过已有的分词工具 VnTokenizer 进行标注。由于分词的效果对后续的诸多环节如词性标注、命名实体和机器翻译等有很大的影响。为了得到更好的分词效果,本文一方面很注重语料的选择,选取了包含政治、经济、文化、体育和娱乐等方面的 25 981 条越南语分词句子,经过人工校对,得到 CRFs 训练语料和测试语料,并进行十倍交叉验证实验。同时选取以上各方面的越南语句子 1 000 条作为比较实验的测试语料。另一方面结合越南语语言的特点,定义了其基本特征,并融入到 CRFs 模型。同时,为了解决交叉歧义给分词带来的困难,基于词典的正向和逆向最大匹配算法抽取了 5 377 条歧义片段,并通过最大熵模型训练成交叉歧义模型,加入到分词模型中,最终实现了越南语分词模型,实验结果证明了本文提出的越南语分词方法的有效性。下一步工作还需要针对歧义分词和未登录词的分词研究更有效的特征选择。

参考文献:

[1] 何国斌,赵晶璐. 基于最大匹配的中文分词概率算法研究[J]. 计算机工程,2010,36(5):173-175.
He Guobin, Zhao Jinglu. Research on probailistic algorithm of Chinese word segmentation based on the maximum match[J]. Computer Engineering,2010,36(5):173-175.

[2] 刘群,张华平. 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.
Liu Qun,Zhang Huaping. Chinese lexical analysis using cascaded hidden markow model[J]. Journal of Computer Research and Development,2004,24(2):1421-1429.

[3] 石民,李斌,陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究[J]. 中文信息学报,2010,24(2):39-45.
Shi Min,Li Bin,Chen Xiaohe. CRF based research on a unified approach to word segmentation and POS tagging for pre-Qin Chinese[J]. Journal of Chinese Information,2010,24(2):39-45.

[4] Phuong L H, Huyen N T M, Azim R, et al. A hybrid approach to word segmentation of vietnamese texts[C]// Proceedings of the 2nd International Conference on Language and Automata. Theory and Applications. Tarragona,Spain: Springer, 2008: 240-249.

[5] 梁远,祝仰修. 现代越南语语法[M]. 广州: 世界图书出版广东有限公司,2012.
Liang Yuan, Zhu Yangxiu. Modern Vietnamese grammar [M]. Guangzhou:World Book Publishing Co., Ltd., 2012.

[6] 阮越雄. 越南语汉源词研究史[D]. 长沙:湖南师范大学,2014.
Nguyen viet hung. Study on the history of Vietnamese Chinese loanwords [D]. Changsha: Hunan Normal University, 2014.

[7] 莫子祺. 从构词方法看越南语同义近义词的用法规律[J]. 学园,2014(28):57-60.
Mo Ziqi . A study on the word-formation methods Vietnamese synonymous synonyms for the usage patterns [J]. Chinese Academy of Sciences, 2014(28): 57-60.

[8] Huihsin T,Pichuan C,Galen A,et al. A conditional random field word segmenter for sighan bakeoff 2005[C]// Proceedings of the fourth SIGHAN workshop. Jeju Isand,Korea:[s. n.],2005:168-172.

[9] 张梅山,邓知龙,车万翔,等. 统计与词典相结合的领域自适应中文分词[J]. 中文信息学报,2012,26(2):8-12.
Zhang Meishan,Deng Zhilong,Che Wanxiang,et al. Combining statistical model and dictionary for domain adaption of Chinese word sementation[J]. Journal of Chinese Information,2012,26(2):8-12.

[10] 刘华明,毕学慧,王维兰,等. 基于最大熵和局部优先度的裂痕唐卡分割[J]. 数据采集与处理,2015,30(2):424-433.
Liu Huaming, Bi Xuehui, Wang Weilan,et al. Crack segmentation based on maximum entropy and local priority[J]. Journal of Data Acquisition and Processing, 2015,30 (2): 424-433.

[11] 汪全全,王靖琰,李勇平. 最大熵矢量量化及其在 TMS320DM642 的实现[J]. 数据采集与处理,2012,27(6):640-645.
Wang Quanquan, Wang Jingyan, Li Yongping. Maximum entropy vector quantization and its implementation in TMS320DM642 [J]. Journal of Data Acquisition and Processing, 2012,27 (6): 640-645.
[12] Della Pietra S, Della Pietra V, Lafferty J. Inducting features of random fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1997,19(4):380-393.
[13] Wallach H. Efficient traning of conditional random fields[EB/OL]. http://www.cogsci.ed.ac.uk,2015-01-14.
[14] 郭剑毅,余正涛,薛征山,等. 基于层叠条件随机场的旅游领域命名实体识别[J]. 中文信息学报,2009,23(5):47-52.
Guo Jianyi, Yu Zhengtao, Xue Zhengshan, et al. Named entity recognition for the tourism domain based on cascaded conditional random fields[J]. Journal of Chinese Information,2009,23(5):47-52.
[15] Sutton C, McCallum A. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data[J]. Journal of Machine Learning Research,2007, 8: 693-723.

作者简介:

熊明明(1987-),男,硕士研究生,研究方向:自然语言处理,E-mail:504609184@qq.com。

李英(1991-),硕士研究生,研究方向:自然语言处理与句法分析, E-mail:1224005374@qq.com。

郭剑毅(1964-),教授,硕士生导师,研究方向:自然语言处理、信息抽取和机器学习等。

毛存礼(1977-),博士,讲师,研究方向:自然语言处理、信息检索。

余正涛(1970-),教授,博士生导师,研究方向:自然语言处理、机器翻译等机器学习。