ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 4,Jul. 2019,pp. 577－584
DOI:10. 16337/j. 1004-9037. 2019. 04. 001

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

融合语言特性的越南语兼类词消歧

郭剑毅[1,2]　赵　晨[1]　刘艳超[1]　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. 昆明理工大学云南省人工智能重点实验室,昆明,650500)

摘　要: 兼类词歧义直接影响词性标注的准确率。本文针对越南语兼类词歧义问题提出一种融合语言特性的越南语兼类词消歧方法。通过构建越南语兼类词词典和兼类词语料库,分析越南语的语言特征和兼类词特点,选取有效的特征集;然后利用条件随机场能添加任意特征等优点,在使用词和词性上下文信息的同时,引入句法成分和指示词特征,得到消歧模型。最后在兼类词语料上实验,准确率达到了87.23%。实验表明本文所提出的越南语兼类词消歧方法有效可行,可以提高词性标注正确率。

关键词: 兼类词消歧;兼类词词典;兼类词语料库;语言特征;条件随机场模型;越南语

中图分类号: TP391　　**文献标志码:** A

Vietnamese Multi-category Words Disambiguation Combined with Language Features

Guo Jianyi[1,2], Zhao Chen[1], Liu Yanchao[1], Mao Cunli[1], Yu Zhengtao[1,2]

(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China; 2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, 650500, China)

Abstract: Multi-category words disambiguation directly affects the part of speech (POS) tagging accuracy. This paper proposed a statistical disambiguation method combined with linguistic characteristics of Vietnamese multi-category words. First, the paper builds Vietnamese multi-category words dictionary and Vietnamese multi-category words corpus, and selects effective feature sets for multi-category words by analyzing of Vietnamese language and multi-category words. Secondly, the paper takes into account the advantages of adding any features of CRFs model, introduces the syntactic and lexical features excepting the features of words and POS, and then builds up the disambiguation model. Finally, testing is carried out on the real multi-category category words corpus, and the accuracy is 87.23%. Experimental results show that the proposed Vietnamese multi-category words disambiguation model is effective and feasible, which can improve the correct rate of POS tagging.

Key words: multi-category words disambiguation; multi-category words dictionary; multi-category words corpus; linguistic characteristics; conditional random fields model; Vietnamese

---

基金项目: 国家自然科学基金(61262041,61562052,61662041)资助项目;国家自然科学基金重点(61732005)资助项目。
收稿日期: 2017-10-31;**修订日期:** 2019-06-28

引　言

兼类词是指一个词具有两个或者两个以上的词性[1]。词性(Part of speech, POS)自动标注是自然语言处理中的基础课题;由于兼类词歧义影响着词性标注的准确率[2],直接影响着词性语料库的质量;而越南语词性语料库的质量是后续语言处理工作的保证,这使它广泛地应用于多个方面,例如:名词短语分析[3]、句法分析[4]和机器翻译[5-6]等。因此,解决越南语兼类词消歧问题是构建高质量的越南语词性语料库的必要条件。

近年来,国内外学者对兼类词消歧方法进行了研究,主要有以下3种:(1)基于规则的方法[5,7-8]。根据北印度语语法,Gupta等[7]提出基于规则的方法,对兼类词进行消歧;Liu等[8]提出基于配置的定量分析现代汉语中动词和名词兼类的分类方法来解决汉语中动-名词兼类问题,根据句法和语义特征对动-名词兼类进行研究;Li等[5]针对中-英专利机器翻译中的动词和介词的兼类,提出基于规则的识别方法,提高了机器翻译质量。(2)基于统计机器学习的方法[9-10]。Dinesh等[9]针对马拉雅拉姆语提出有监督语言模型,同时该模型引入命名实体识别器和词法分析器,进行兼类词消歧;针对电子商业领域的兼类词,Fei F等[10]提出了基于条件随机场消歧方法,减少汉语中电子商业的歧义,同时提高了用户检索体验。(3)基于混合的方法[11-12]。Zhang等[11]对汉语中的兼类词采用集成模型进行词性消歧,准确率达到89.69%;Xia等[12]针对汉语提出基于规则和统计的方法进行兼类词消歧,使用多种统计方法进行消歧,对消歧结果中不理想的兼类词采用规则方法再次进行消歧,以上的研究都已取得较好的结果。

上述研究主要针对英语、汉语等语言,就越南语兼类词消歧而言,相关研究相对较少。兼类词歧义消歧属于词性标注范畴,在越南语词性标注方面,文献[13]在支持向量机(Support vector machine, SVM)模型中融入普通特征(词汇特征、词的上下文特征、词性特征和拼写特征)和特殊特征(重复特征、前缀和后缀特征),进行词性标注,正确率为93.51%;文献[14]将词特征和音节特征融合到统计模型SVM、最大熵模型(Maxinum entropy model, MEM)和条件随机场(Conditional random fields, CRFs)中建模并进行分词,比较3种模型的结果;文献[15]提出了最大熵方法融入基本特征和音节特征,正确率达到93.40%,但这些研究几乎没有考虑兼类词问题。

目前,随着中越两国文化和经济交流的日益频繁,汉越自然语言处理越来越重要,越南语兼类词消歧工作迫在眉睫。但越南语兼类词消歧研究工作很少,为了提高越南语词性标注质量,本文通过分析越南语的语言和兼类词特点,提出了融合语言特性的越南语兼类词消歧方法。

1　兼类词消歧框架

借鉴已有的兼类词消歧方法和思路,本文提出的方法原理框架如图1所示,主要包括:越南语语料预处理、构建越南语兼类词字段和越南语兼类词词典、构建基于条件随机场的消歧模型和语料测试等过程。

图1中,越南语兼类词消歧的具体流程如下:(1)越南语语料预处理。本文从越南语网站中抽取具有政治、文化、经济

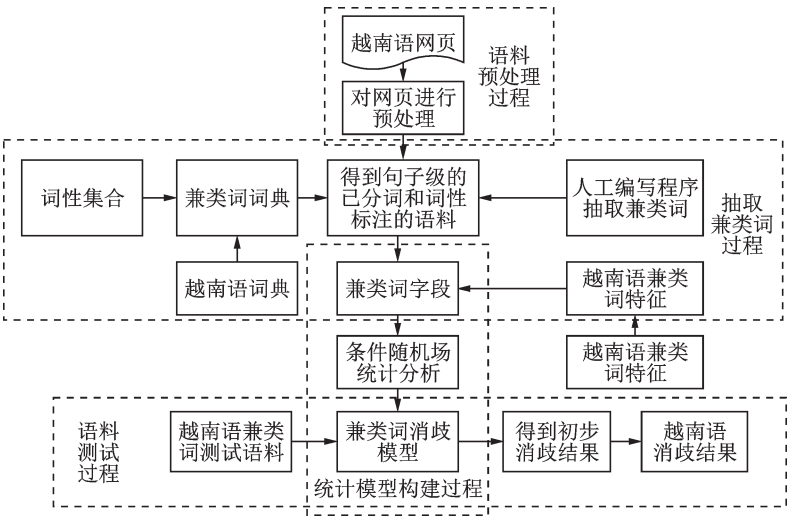

图1　越南语兼类词消歧框架图

Fig.1　Vietnamese multi-category words disambiguation framework

和新闻等类型题材的网页,通过爬虫程序,获得越南语文本语料;对其进行去噪等操作,使用分词工具进行分词,使用词性标注工具进行词性标记,并完成校对;(2)构建字段语料库和越南语兼类词词典。通过人工对越南语字典分类整理得到1 659条的兼类词词典;以此为基础,从已构建的词性标注语料库中通过编程抽取396 946条越南语兼类词字段语料;(3)构建基于条件随机场的消歧模型。根据越南语中兼类词的特点,选取消歧特征,将其与已抽取的越南语兼类词字段向融合,形成训练语料,使用条件随机场模型进行建模,获得基于条件随机场的消歧模型;(4)语料测试。用构建的基于条件随机场的消歧模型对测试语料进行消歧,得到消歧结果。

2 越南语兼类词消歧模型

2.1 越南语及其兼类词特点

越南语属于南亚语系,和汉语一样是孤立语,但其由拉丁字母、表音文字及标点符号等构成。越南语的主要特点如下:(1)由一个或多个词素构成;(2)修饰语位于被修饰词之后;(3)越南语由于受多元文化的影响,在书写及表达方式上显示出复杂性和多样性。越南语兼类词的特点主要有:(1)大多数的越南语兼类词都是常用词,主要集中在名词、动词、形容词和量词(单位词)等词性之间的转化上,如,thuốc men(药品、用药,名词兼动词);又如,bát(碗,名词兼量词),một cái bát(一个碗),một bát cơm(一碗饭);(2)在某些词前出现其他词时,这些词的词性会发生转变,例如动词前有"sự,(事),việc(事情)..."时,该动词会变成名词使用;(3)词的语义搭配关系存在一定的优先关系。兼类词消歧工作的复杂程度一般随着词性标注集划分规模程度来决定,一般来说,越是常用的词其词义活用的现象越严重,词的兼类情况就越复杂。造成越南语兼类词现象的主要原因有:(1)吸收外来文化;(2)词义的派生;(3)越南语词的活用等,以上现象给越南语兼类词消歧工作带来困难和挑战。

2.2 统计消歧模型

从上分析可知,越南语兼类词消歧需要结合越南兼类词和语言的结构特点。与传统消歧模型相比,条件随机场模型具备融合不同特征的功能,能够使用复杂、有重叠性和非独立性的特征进行训练和推理,能够充分利用上下文信息和其他外部信息作为特征;同时能适当地避免数据标注偏执问题和歧义问题。因此本文选取条件随机场建立消歧模型。

2.2.1 条件随机场原理

CRFs是由John Lafferty等提出的一种统计机器学习模型,它结合了最大熵模型和隐马尔科夫模型的特点,近年来在分词、POS标签和名词组块识别等序列标注任务中取得了很好的效果。它是一种无向图模型,在待标注的观测序列确定的情况下,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[1]。对于给定的一个长度为n的序列,$X=x_1,x_2,x_3,\cdots,x_n$,则输出$Y=y_1,y_2,y_3,\cdots,y_n$的概率可以定义为

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

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

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

式中:$Z(x)$为一个归一化因子,$f_k\left(y_t,y_{t-1},x_t\right)$是对整个序列的$X$标记位于$t$和$t-1$位置上标记的转移概率$\lambda_k$是每一个特征特征权重向量。

在本文的越南语兼类词消歧模型中,条件随机场通过训练语料得到模型参数的最优值,使消歧结果最优化。

2.2.2 特征选取

条件随机场模型的性能取决于特征的选取。根据越南语的语言特性和兼类词特征,本文主要选取以下4种特征,其特征模板如表1所示。其中w表示词,p表示词性,g表示句法成分,Pz/Sz表示指示词特征,具体含义在实验结果与分析中进行说明。

表1 特征模板

Tab. 1 Feature template

特征	特征符号	特征含义
词及上下文信息	w(−2),w(−1),w(0),w(1),w(2); w(−2)/w(−1),w(−1)/w(0),w(0)/w(1),w(1)/w(2); w(−2)/w(−1)/w(0),w(−1)/w(0)/w(1),w(0)/w(1)/w(2)	表示选取当前词与上下文信息作为有效特征
词性及上下文信息	p(−2),p(−1),p(1),p(2); p(−2)/p(−1),p(1)/p(2);	表示选取当前词词性周围信息作为有效特征
句法成分特征	g(−1),g(0),g(1) g(−1)/g(0),g(0)/g(1),g(−1)/g(0)/g(1)	关系搭配特征及上下文
指示词特征	Pz/Sz	Pz表示前指示词/Sz后指示词,若无,表示NULL

(1)词特征。由于词形态的改变能表征词以及其含义的改变,上下文的词能当前词产生影响。例如“cuốc”词在与“đất”搭配时,词性为动词,在与“cái”搭配时,词性为名词。因此,本文选取词以及上下文信息做为有效特征。

(2)词性特征。兼类词的词性会受到其前后两个词的词性的影响,如“bát(碗)”有量(单位)词和名词两种词性,在“một bát cơm(一碗饭)”中,由于“một(一)”是数词,“cơm”是名词,从而可以判断“bát(碗)”是量词。因此,本文选取词性以及上下文词性信息作为有效特征。

(3)句法成分特征。在越南语中,语义搭配关系符合一定规律。例如越南语句子结构一般为“主-谓-宾”,兼类词作为宾语成分接在动词后面时,一般为名词词性;越南语中,被修饰语的词性,可以通过位于被修饰词后面的修饰语来确定;副词或者形容词前面一般搭配谓语,被修饰词应该首先优先考虑动词等等,如果无成分特征,则表示NULL,否则表示主语(S)、谓语(V)和宾语(O)等。因此,本文选取当前兼类词所充当句法成分和周围成分特征作为有效特征。

(4)指示词特征。在越南语中,一些特定的指示词出现在某些词前时,这些词的词性会发生转变,例如“sự(事),việc(事情),cuộc(量词)”等词出现在动词前时,该动词会变为名词词性;“một(一)”等词出现在名词前时,该名词会变为量词词性。在越南语中,前指示词和后指示词特征需要考虑;如果缺失指示词,表示为NULL。因此,本文选取指示词作为有效特征。

3 实验分析

3.1 实验评价标准

为了评估本文方法的消歧效果,实验将采用中英文消歧常采用的评价标准:准确率(Precision)(正确消歧越南语兼类词个数与消歧兼类词总数的比值)来作为本文评价标准。

$$\text{准确率(Precision)}=\frac{\text{正确消歧兼类词个数}}{\text{消歧兼类词总数}}\times 100\% \tag{3}$$

式中准确率数值在0和1之间,越接近1,就表明本文的方法越有效。

3.2 实验数据

本文实验所用实验数据包括兼类词词典和兼类词字段语料。目前,由于越南语兼类词的相关研究资源匮乏,故本文需要构建语料库。兼类词词典是由越南语字典经过本文人工处理所得到的,包含1 659个兼类词;越南语兼类词字段库是本文通过编写程序对越南语文本语料,经过抽取得到的(包括新闻、政治、经济等方面),共有396 946条兼类词字段信息,所有字段保存为"UTF-8"格式,在本文实验语料中不存在未登录兼类词,其采用的词性标注集是文献[16]制定的词性集合(19种类型)。其中分词方法使用文献[17]中的方法,准确率在96.86%。抽取兼类词字段流程图如图2所示。

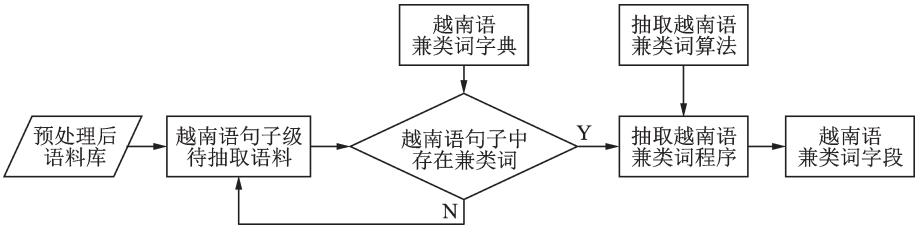

图2 越南语兼类词抽取流程图

Fig.2 Vietnamese multi-category words extraction flowchart

越南语兼类词抽取算法如下:

输 入:预处理后的语料库。

第1步:从预处理后的语料中抽取1条越南语句子级语料,执行第2步;

第2步:根据越南语兼类词词典,判断获取到的句子级语料中是否含有兼类词出现,执行第3步;

第3步:如果句子级语料中存在兼类词,用程序抽取兼类词,执行第4步;否则,执行第1步;

第4步:抽取得到兼类词字段,返回第1步。

输 出:越南语兼类词字段。

3.3 实验结果与分析

本文实验的实验语料选用的是3.2节中得到的396 946条兼类词字段语料,除实验2外所用的语料是将所有语料分为5份,选用其中4份用于训练,另外1份用于测试。具体格式如表2所示。

表2 兼类词字段语料具体格式

Tab. 2 Specific format of the multi-category words corpus

越南语	词性	指示词	句法	词性
tham_quan	V	O	−1_root	V
di_tích	N	O	−1_dob	N
lịch_sử	O	O	−1_nmod	N
,	O	O	−3_punct	CH
văn_hóa	N	O	−4_sub	N
trên	O	O	−4_loc	E
địa_bàn	N	O	−1_pob	N
Thành_phố	O	O	−1_nmod	N
Hồ_Chí_Minh	N	O	−2_nmod	N
;	O	O	−9_punct	CH

实验中的条件随机场模型使用CRF^{++}工具包实现，其中template文件中的内容由2.2.2节中的特征模板得到，其中w(−2)转为%x[−2,0]，w(−2)/w(−1)转为%x[−2,0]/ %x[−1,0]，w(−2)/w(−1)/w(0)转为%x[−2,0]/ %x[−1,0]/ %x[0,0]以此类推，p(−2)转为%x[−2,1]，p(−2)/p(−1)转为%x[−2,1]/ %x[−1,1]，g(−1)转为%x[−2,3]，g(−1)/g(0)转为%x[−2,3]/ %x[−1,3]，Pz转为%x[−1,2]，Sz转为%x[1,2]。为了验证本文方法的有效性，本文从不同角度设计以下3组实验：

实验1 为了考察4类特征对越南语兼类词消歧统模型的贡献度，分别将4类特征单独融入消歧模型中，特征的贡献程度通过准确率进行比较，实验结果如表3，图3所示。

表3 4类特征对模型贡献度实验

Tab. 3 Model contribution test of four types of characteristics

类型	准确率/%
词特征	69.70
词性特征	63.81
句法成分特征	60.30
指示词特征	57.10
融合4种特征	87.80

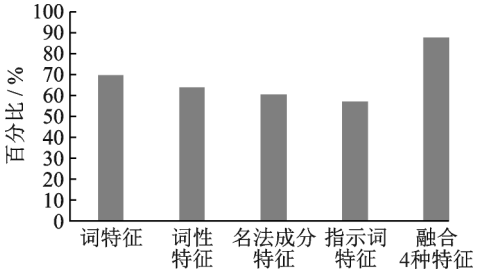

图3 4类特征对模型贡献度实验

Fig.3 Model contribution test of four types of characteristics

从图3中可以看出，单独使用词特征的准确率为69.70%，比单独使用词性特征高5.89%，其主要原因为单独使用词性特征时，可能会造成词性搭配上的歧义，而单独使用词特征时，词性是确定的；句法成分特征相较于前两个特征偏低，其主要原因为在不同词性表示相同的句子成分时，区分度不强造成；指示词特征正确率最低，这说明并非所有兼类词都有指示词特征，能通过指示词消歧的兼类词较少；融入所有特征后的模型准确率最高。由此可见，词特征和词性特征较为有效，然后是句法成分特征和指示词特征。

实验2 为了评估所提出的条件随机场统计模型的效果，将396 946条兼类词字段平均分为5份，选取其中1份作为测试语料，其他4份作为训练语料，进行5倍交叉验证实验，求其平均准确率，作为条件随机场模型兼类词消歧的测评结果，实验结果如表4，图4所示。从表4，图4可以看出，序号1的实验准确率达到88.15%，达到了局部最优。实验平均准确率为87.23%，作为所提出的条件随机场统计模型的效果。

实验3 最大熵建模和支持向量机是自然语言处理中常用到的模型[16,18]，最大熵只需要集中精力选择特征，而不需要花费精力考虑如何使用这些特征；同时该模型不需要像其他模型中常常使用的独立性假设，而支持向量机在小样本训练集上能够得到很好的结果，且具有优秀的泛化能力是效果最好

表4 5倍交叉验证实验

Tab. 4 Five times cross validation experiment

序号	语料分配	准确率/%
1	第1份作为测试语料，其他4份作为训练语料	88.15
2	第2份作为测试语料，其他4份作为训练语料	87.98
3	第3份作为测试语料，其他4份作为训练语料	86.14
4	第4份作为测试语料，其他4份作为训练语料	86.35
5	第5份作为测试语料，其他4份作为训练语料	87.54
平均值		87.23

的分类器之一。故本文选这两个模型和条件随机场模型进行比较。本实验采用相同的特征、训练语料和测试语料,分别使用这3种模型进行实验,模型性能通过准确率进行对比,实验结果如图5,表5所示,其中“1”代表“词特征”类型,“2”代表“词性特征”类型,“3”代表“句法成分特征”类型,“4”代表“指示词特征”类型。

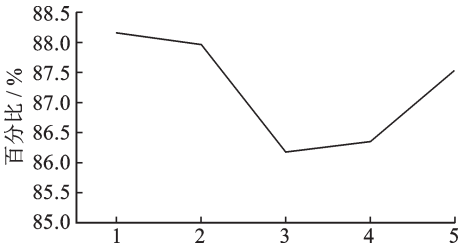

图4 5倍交叉验证实验

Fig.4 Five times cross validation experiment

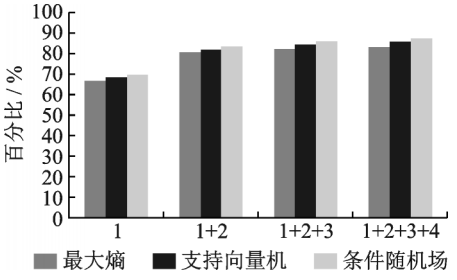

图5 不同模型比较

Fig.5 Comparison of different models

从图5,表5可以看出,在使用相同特征时,条件随机场模型比最大熵和支持向量机效果好,可见,条件随机场模型能和本文的特征更好的融合;在词特征的基础上,融入词性特征,准确率提高13.73%;在词和词性特征基础上,融入句法成分特征时,模型准确率提高了2.17%;融入所有特征,模型整体性能有所提高,该模型准确率到达了87.23%。由此可见,本文所提出的基于条件随机场的越南语兼类词消歧方法有效可行。

表5 不同模型比较

Tab. 5 Comparison of different models

方 法	特征	准确率/%
最大熵	1	66.51
最大熵	1+2	80.68
最大熵	1+2+3	82.15
最大熵	1+2+3+4	82.87
支持向量机	1	68.51
支持向量机	1+2	81.92
支持向量机	1+2+3	84.21
支持向量机	1+2+3+4	85.74
条件随机场	1	69.70
条件随机场	1+2	83.43
条件随机场	1+2+3	85.60
条件随机场(本文的方法)	1+2+3+4	87.23

4 结束语

兼类词消歧直接影响着词性标注的准确率。本文针对越南语兼类词歧义问题,提出了一种融合语言特性的越南语兼类词消歧方法。通过构建越南语兼类词词典和字段语料库,分析越南语的语言特性和兼类词特征,选取了词特征、词性特征、句法成分特征和指示词特征这4种有效特征,采用条件随机场进行建模,得到越南语兼类词的统计消歧模型,在真实语料库上,实验获得了良好的效果。实验结果表明,本文所提出的融合语言特性的越南语兼类词消歧方法能有效解决越南语兼类词歧义问题。本文将不断补充语料、挖掘更多的越南语语言特征和兼类词特点,尝试新方法进行越南语兼类词消歧,进一步提高兼类词的消歧性能。

参 考 文 献:

[1] 刘艳超. 越南语浅层句法分析方法的研究[D]. 昆明:昆明理工大学, 2017.
Liu Yanchao. A study on the method of shallow syntactic analysis in Vietnamese[D]. Kunming: Kunming University of Science and Technology, 2017.

[2] Qian Y L, Zheng J H. An approach to improving the quality of part-of-speech tagging of Chinese text[C]//International Conference on Information Technology: Coding and Computing. [S.l.]: ITCC, 2004: 183.

[3] Thao N T H, Thai N P, Minh N L, et al. Vietnamese noun phrase chunking based on conditional random fields[C]// Knowledge and Systems Engineering, International Conference on. [S.l.]: IEEE, 2009: 172-178.

[4] Nguyen P T, Le A C, Ho T B, et al. Vietnamese treebank construction and entropy-based error detection[J]. Language Resources & Evaluation, 2015, 49(3): 487-519.

[5] Li H, Zhu Y, Jin Y. Identifying verb-preposition multi-category words in Chinese-English patent machine translation[C]// Australasian Conference on Artificial Life and Computational Intelligence. Australia: Springer, 2015: 409-421.

[6] Xiong J, Zhong L, Wang A. Example and ontology based machine translation for oracle bone inscriptions[J]. Journal of Huazhong University of Science & Technology, 2013, 41(S2): 222-226.

[7] Gupta J P, Tayal D K, Gupta A. A TENGRAM method based part-of-speech tagging of multi-category words in Hindi language[J]. Expert Systems with Applications, 2011, 38(12): 15084-15093.

[8] Liu P, Ding J. A research into the multi-category words of verb-noun in modern Chinese based on the quantitative analysis of the collocating classifiers[C]//Workshop on Chinese Lexical Semantics. [S.l.]: Springer International Publishing, 2015: 294-306.

[9] Dinesh T, Jayan V, Bhadran V K. Word category disambiguation for malayalam: A language model approach[C]// International Conference on Computational Science, Engineering and Information Technology. [S.l.]: ACM, 2012: 642-646.

[10] Fei F, Yang Y, Xu W, et al. An effective resolution method of Chinese multi-category words with conditional random field in electronic commerce[C]//International Conference on Neural Information Processing. [S.l.]: Springer, 2015: 562-570.

[11] Zhang Y, Qu W, Liu J, et al. Research on disambiguation of multiple syntactic category words based on ensemble of classifiers [J]. Journal of Nanjing Normal University, 2010, 33(4): 144-143.

[12] Xia J, Chai Y M, Zan H Y. Study on multi-category of common words based on statistics and rules[J]. Computer Engineering and Design, 2013, 34(2): 654-659.

[13] Nghiem M, Dien D, Nguyen L M. Improving Vietnamese POS tagging by integrating a rich feature set and support vector machines[C]//Proceedings of Research, Innovation and, Vision for the Future. [S.l.]: IEEE, 2008: 128-133.

[14] Oanh Thi Tran, Cuong Anh Le, Thuy Quang Ha, et al. An experimental study on vietnamese POS tagging[C]//Proceedings of the International Conference on Asian Language Processing. Singapore: IALP, 2009: 23-27.

[15] Phuong Le-Hong, Azim Roussanaly, Thi Minh Huyen Nguyen, et al. An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts[C]//Proceedings of TALN 2010. Montreal, Canada: [s.n.], 2010: 19-23.

[16] 熊明明. 越南语词法分析方法研究[D]. 昆明:昆明理工大学, 2016.
Xiong Mingming. A study of the method of analysis of the Vietnamese language[D]. Kunming: Kunming University of Science and Technology, 2016.

[17] 熊明明, 李英, 郭剑毅, 等. 基于CRFs和歧义模型的越南语分词[J]. 数据采集与处理, 2017, 32(3): 636-642.
Xiong Mingming, Li Ying, Guo Jianyi, et al. Vietnamese word segmentation with conditional random fields and ambiguity model[J]. Journal of Data Acquisition and Processing, 2017, 32(3): 636-642.

[18] 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 a1. Segmentation of RIP Tangka based on maximum entropy and local priority[J]. Journal of Data Acquisition and Processing, 2015, 30(2): 424-433.

作者简介:

郭剑毅(1964-),女,教授,研究方向:自然语言处理、信息抽取以及知识获取。

赵晨(1993-),男,硕士研究生,研究方向:自然语言处理,E-mail:619235831@qq.com。

刘艳超(1990-),男,硕士研究生,研究方向:自然语言处理。

毛存礼(1977-)男,博士,副教授,研究方向:自然语言处理、信息检索、机器翻译以及智能决策系统。

余正涛(1970-),男,博士,教授,研究方向:自然语言处理、信息检索及机器翻译,E-mail:ztyu@hotmail.com。

(编辑:刘彦东)