Journal of Computer Applications
计算机应用,2014,34(8):2197 -2201
ISSN 1001-9081
CODEN JYIIDU
2014-08-10
http://www.joca.cn

文章编号:1001-9081(2014)08-2197-05
doi:10.11772/j.issn.1001-9081.2014.08.2197

基于规则的汉语兼类词标注方法

李华栋,贾　真*,尹红风,杨　燕
(西南交通大学 信息科学技术学院,成都 610031)
(*通信作者电子邮箱 zjia@swjtu.cn)

摘　要:针对目前汉语兼类词标注的准确率不高的问题,提出了规则与统计模型相结合的兼类词标注方法。首先,利用隐马尔可夫、最大熵和条件随机场3种统计模型进行兼类词标注;然后,将改进的互信息算法应用到词性(POS)标注规则的获取上,通过计算目标词前后词单元与目标词的相关性获得词性标注规则;最后,将获取的规则与基于统计模型的词性标注算法结合起来进行兼类词标注。实验结果表明加入规则算法之后,平均词性标注准确率提升了5%左右。

关键词:词性标注;互信息;汉语兼类词;规则;中文信息处理
中图分类号:TP391.1　　**文献标志码**:A

Rule-based tagging method of Chinese ambiguity words

LI Huadong, JIA Zhen*, YIN Hongfeng, YANG Yan
(*School of Information Science and Technology, Southwest Jiaotong University, Chengdu Sichuan* 610031, *China*)

Abstract: Concerning the low accuracy of tagging Chinese ambiguity words, a combined tagging method of rules and statistical model was proposed in this paper. Firstly, three kinds of traditional statistical models, including Hidden Markov Model (HMM), Maximum Entropy (ME) and Condition Random Field (CRF), were used to tagging problem of the ambiguity words. Then, the improved mutual information algorithm was applied to learn Part Of Speech (POS) tagging rules. Tagging rules were got through the calculation of correlation between the target words and the nearby word units. Finally, rules were combined with statistical model algorithm to tag Chinese ambiguity words. The experimental results show that after adding the rule algorithm, the average accuracy of POS tagging promotes by 5%.

Key words: Part Of Speech (POS) tagging; mutual information; Chinese ambiguity word; rule; Chinese information processing

0　引言

词性标注作为自然语言处理的基础研究内容之一,多年来一直是研究的热点。兼类词标注的准确率是影响词性标注准确率的主要因素之一,兼类词问题本质上是在对目标词进行词性标注时的歧义问题。针对这个问题,国内外学者进行了大量研究,而这些研究主要是集中在对兼类词的特征描写上,而很少有人对研究兼类词的自动识别问题进行研究。词的兼类现象是汉语中一个很常见又很复杂的问题,由于人们对汉语词类认识的匮乏和有限,汉语词的兼类问题仍然存在很多的争议,不同的语言学家对汉语的理解不同,对中文词类的划分也不同。目前,汉语词类的划分没有统一的标准,比较典型的有清华大学的《汉语树库》词性标记集和《人民日报》词性标记集等。

兼类词的特点是数量少,但是使用的频率非常高。本文对1998年1月份《人民日报》语料库进行统计,发现总共有5477个兼类词,占总词数的7.3%,而这些兼类词总共出现了105033次。由此可见,对这些兼类词的词性进行正确的标注对总体的词性标注准确率至关重要。

目前对于兼类词的词性标注的研究,有许多相关的技术和文献:早期基于规则的方法使用的词性标注规则是由语言学家根据语言规律进行人工书写完成的,这使得规则的编写不仅费时费力,而且容易出现规则冲突、规则不完备等问题[1-2],因而势必影响到标注正确率。针对人工编写规则带来的问题,近10多年来出现了很多新的词性标注方法,它们都是从真实语料中利用机器学习原理获取消除词性兼类歧义的语言学知识。根据词性标注知识描述方式不同,这些基于语料库的方法大致可以分为两类:一类是规则学习方法,即从真实语料中自动获取词性标注规则,如基于转换的错误驱动(Transform-Based Error Driven, TBED)方法[3]、基于决策树(Decision Tree,DT)方法[4]等;另一类是统计方法,即用某种统计模型作为词性标注知识的描述方式,如隐马尔可夫模型(Hidden Markov Model, HMM)方法[5]、最大熵(Maximum Entropy,ME)方法[6]、条件随机场(Condition Random Field, CRF)方法[7]等。

本文先采用统计算法对兼类词进行标注,针对传统手工编辑词性标注规则库效率和准确率过低的问题,提出了一种利用互信息算法获得汉语词性标注规则的方法,通过计算目

收稿日期:2014-04-06;**修回日期**:2014-05-06。　**基金项目**:国家自然科学基金资助项目(61134002,61170111,61202043,61262058)。
作者简介:李华栋(1988-),男,湖北黄冈人,硕士研究生,主要研究方向:自然语言处理、数据挖掘;　贾真(1975-),女,河南开封人,讲师,主要研究方向:信息抽取、知识工程;　尹红风(1963-),男,河南夏邑人,教授,主要研究方向:语义搜索、大数据;　杨燕(1964-),女,四川成都人,教授,CCF会员,主要研究方向:数据挖掘、计算智能、集成学习。

标词与目标词前后的词单元之间的相关性来获取词性标注规则，并且将所获得的规则与统计模型相结合进行词性标注。这种方法克服了传统的规则学习方法迭代次数过多的问题以及人工编写规则覆盖率不高而且效率低的问题。实验结果表明，这种方法能够明显提高兼类词标注的准确率。

1 兼类词词性标注模型

词性标注是自然语言处理领域的基础，可以提高信息检索的效果和效率，它在信息检索领域有着非常重要的作用。国内外该方面研究人员很重视它，成功设计出很多词性标注模型[8-13]。本文所使用的词性标注模型如图 1 所示。

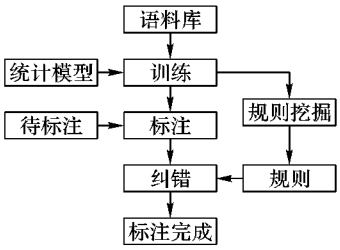

图 1 词性标注模型

整个词性标注模型中有两个重要的因素：

1）建立统计模型。训练统计模型，对句子进行初步的词性标注。

2）规则挖掘。从已标注的语料库中找出目标词的词性与其前后的词和前后词的词性之间的关联，将那些明显相关的词和词性作为本文的词性标注规则。利用挖掘出来的规则对基于统计模型的词性标注结果进行纠错。

2 建立统计模型

本文所选用的统计模型为隐马尔可夫模型、条件随机场模型和最大熵模型。统计学习主要分为训练和标注两个过程。

实验过程中建立统计模型所使用的语料库是 1998 年 1 月份的《人民日报》和《语料库在线》两种语料库，这两种语料库都是已经分词并且人工标注过的语料库。

2.1 HMM

HMM 为实际应用最广泛的模型，对于句子 $W = w_1w_2\cdots w_{i-1}w_i\cdots w_m$（$w_i$ 为第 i 个单词），可以用 $P(w_1w_2\cdots w_i\cdots w_m)$ 来表示它出现的概率：

$$P(w_1w_2\cdots w_i\cdots w_m) = P(w_1)P(w_2 \mid w_1)\cdots \cdot P(w_m \mid w_1w_2\cdots w_{m-1}) \quad (1)$$

若第 i 个词出现的概率依赖于它前面的 $N-1$ 个词，该模型为 N 元文法模型。将 W 视作一阶 Markov 链，则有二元文法模型（Bigram 模型）：w_i 单词出现的概率 $P(w_i \mid w_1w_2\cdots w_{i-1})$ 只依赖于 w_{i-1}。本文使用的是一阶 Markov 链。HMM 的解码就是求使 $P(w_1w_2\cdots w_i\cdots w_m)$ 值最大的过程，算法采用 Viterbi 算法解码。

2.2 ME 模型

最大熵原理原本是热力学中一个非常重要的原理，后来被广泛应用于自然语言处理方面。其基本原理很简单：对所有的已知事实建模，对未知不作任何假设。

若将词性标注或者其他自然语言处理任务看作一个随机过程，最大熵模型就是从所有符合条件的分布中，选择最均匀的分布，此时熵值最大。求解最大熵模型，可以采用拉格朗日乘数法，其计算公式为：

$$p_\lambda(y \mid x) = \frac{1}{Z_\lambda(x)}\exp\Big[\sum_i \lambda_i f_i(x,y)\Big] \quad (2)$$

其中：$Z_\lambda(x) = \sum_y \exp\Big[\sum_i \lambda_i f_i(x,y)\Big]$，是对应特征的权重，$f_i$ 表示一个特征。每个特征对词性选择的影响大小由 λ_i 特征权重决定，而这些权值可由 GIS（Generalized Iterative Scaling）学习算法自动得到。

最大熵模型的关键在于特征选取，特征选取的恰当与否会对结果有直接影响。

为了降低数据稀疏和语料中的一些噪声带来的预测不可靠性，需要对特征进行筛选。特征的筛选方法包括基于频数阈值的方法和增量式特征选择方法。基于频数的特征选择方法主要基于这样一个假设：不常出现的特征是噪声或不相关的，只有那些出现频数比较大的特征才真正代表了数据的特性。这里采用的就是基于频数的方法。本文取前后阈值的数量为 4，这个阈值的大小和训练语料的大小有关，需要从实验中得知。

2.3 CRF 模型

CRF 模型的基本原理定义为对于输入数据序列 x 和标注结果序列 y，条件随机场的全局特征表示为：

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

其中：i 遍历输入数据序列的所有位置，$f(y,x,i)$ 表示在 i 位置时各个特征组成的特征向量。于是，CRF 定义的条件概率分布为：

$$p_\lambda(y \mid x) = \frac{\exp[\lambda \cdot F(y,x)]}{Z_\lambda(x)} \quad (4)$$

其中 $Z_\lambda(x) = \sum_y \exp[\lambda \cdot F(y,x)]$。

给定一个输入数据序列 x，标注的目标就是找出其对应的最可能的标注结果序列 y，即

$$\bar{y} = \arg\max_y p_\lambda(y \mid x) \quad (5)$$

由于 $Z_\lambda(x)$ 不依赖于 y，因此有 $\bar{y} = \arg\max_y p_\lambda(y \mid x) = \arg\max_y \lambda \cdot F(y,x)$。与隐马尔可夫模型相似，CRF 使用 Viterbi 算法来得到最佳的标注结果序列。

CRF 能够同时使用中心词的前 n 个词和后 m 个词作为该词的上下文信息。这样，中心词的词性不仅与它前面的词有关，还与它后面的词有关，更加符合实际情况。在本文中，使用了中心词前后 4 个词和这些词的词性作为特征，每个特征的权值都设为 1。

3 词性标注规则挖掘

3.1 规则挖掘简介

规则挖掘就是寻找目标词的词性与其前后词之间相关性的过程。如果一个词单元 A 与另外一个词单元 B 的某一个词性明显相关，就可以将 A 定义为词 B 的一条词性标注规则。例如："发展"是一个兼类词，"舞剧/n 的/u 发展/vn 成

果/n　和/c　动向/n"中"发展"是动名词,很明显,"的/u"和"发展/vn"有很强的相关性,故可以把"的/u"定义为目标词"发展/vn"的一条词性标注规则,即如果"发展"前面出现"的/u"这个词单元,就将"发展"标注成"vn"。

3.2　规则挖掘方法

为获得两个词单元的相关性,通常要对两个词单元的相关性进行计算,计算的公式如下:

$$MI(A,B) = \frac{p(A,B)}{p(A) \times p(B)} \tag{6}$$

其中:$p(A,B)$ 表示 A 和 B 共现的次数,$p(A)$ 表示训练语料中词单元 A 出现的次数,$p(B)$ 表示训练语料中词单元 B 出现的次数。如果互信息值很小,可以认为 A、B 不相关;如果互信息值很大,却不能说明 A、B 很相关。因为,式(6) 的值依赖于分母中两个单个词单元的频度。例如:A 和 B 是两个很少出现的词,假设 A 和 B 在训练语料中分别只出现了一次,此时 A 和 B 的互信息值就为 1,显然不能判定 A 和 B 很相关。

对于目标词的词性标注规则来说,本文只关心目标词与目标词附近的词单元的相关性,所以要对式(6) 进行改进,改进后的公式为:

$$MI_i(A,B) = p(A,B)/p_{B_i}(A) \tag{7}$$

其中:$MI_i(A,B)$ 表示 B 前面第 i 个位置的词单元和词单元 B 的相关性,$p_{B_i}(A)$ 表示词单元 B 之前的第 i 个位置上词单元 A 出现的次数。可以通过调整 i 的值来调整词性标注规则的窗口大小,在这里,本文设 i 的值为 2,即只关心目标词前后各两个词与目标词之间的相关性。

在《人民日报》语料库中,有些词出现的次数较少,这些出现较少的词不具有说明性,如果目标词 B 之前的第 i 个位置单元词 A 出现的次数较少,本文就认为 A 不是词单元 B 的标注规则。本文方法设阈值为 20,即从获取的词性标注规则中除去单元词 A 出现的次数少于 20 次的那些规则。

例如:要统计"舞剧/n　的/u　发展/vn　成果/n　和/c　动向/n"中"发展"的词性标注规则,本文只关心"舞剧/n""的/u""成果/n"和"动向/n"这几个词和"发展/vn"之间的共现关系,如果要计算"的/u"和"发展/vn"之间的共现概率,本文只关心"发展/vn"的前一个位置"的/u"出现的次数和"发展"前出现"的/u"并且"发展"被标注成"vn"的次数之间的关系。若语料库中"发展"前面出现 100 次"的/u","发展"前出现"的/u"并且"发展"被标注成"vn"的次数为 95 次以上,本文就基本认定"发展"的前一位置为"的/vn"是发展被标注成"/vn"的一条规则。

兼类词的词性规则可以形式化描述为:

$\langle ID\rangle \rightarrow [L_2],[L_1],[R_1],[R_2]$

$L_2 \rightarrow \langle$词 1$\rangle|\langle$词 2$\rangle|\cdots|$a|v|n|j$\cdots$

$L_1 \rightarrow \langle$词 1$\rangle|\langle$词 2$\rangle|\cdots|$a|v|n|j$\cdots$

$R_1 \rightarrow \langle$词 1$\rangle|\langle$词 2$\rangle|\cdots|$a|v|n|j$\cdots$

$R_2 \rightarrow \langle$词 1$\rangle|\langle$词 2$\rangle|\cdots|$a|v|n|j$\cdots$

其中:ID 为所识别的兼类词的词性,L_2 表示目标词左边第 2 个位置的词语或词性信息,L_1 表示目标词左边第 1 个位置词语或词性信息,R_1 表示目标词右边第 1 个位置的词语或词性信息,R_2 表示目标词右边第 2 个位置的词语或词性信息。

以下是兼类词"工作"的规则描述样例:

MB 工作

$@\langle a\rangle \rightarrow L_2\hat{}L_2 \rightarrow$[做好|把$\cdots$|w]$\rightarrow$vn

$@\langle b\rangle \rightarrow L_1\hat{}L_1 \rightarrow$[的|各项|救灾$\cdots$]$\rightarrow$vn

…

其中:MB 表示目标词;a 和 b 表示规则的标号。

规则$@\langle b\rangle \rightarrow L_1\hat{}L_1 \rightarrow$[的|各项|救灾$\cdots$]$\rightarrow$vn 的解释为:如果"工作"的前一位置出现"的""各项""外交"和"救灾"等这些词的时候,发展就可以被标注成"vn",即动名词。

兼类词的每一个词性规则都可以看成是一个模式表达式,因为符号的特殊性,这个模式语言的定义并不能认为是正则表达式。本文考虑把兼类词的识别问题看成是字符串的匹配问题,而正则表达式在文本字符的处理方面具有高效、易用的优点,考虑将规则中定义的词性进行实例化,然后用实例化后所得到的词集去替换对应的词性字符,再对其他的匹配字符也作相应的转化,就得到了规则的正则表达式;最后对语料在特征属性匹配器上进行字符串的匹配,根据匹配结果确定兼类词的词性编码。

算法的具体步骤如下:

1)读取训练语料,找出含有目标词的那些句子,并存入内存。

2)根据式(7)构建目标词 B 的词性标注规则,并写入文件。

构建规则的时候设阈值为 0.95,即将阈值大于 0.95 的那些条目存入文件,文件格式如下:@A:位置 @B:词 @C:词性 @D:概率,第 1 列代表词的位置,第 2 列代表标注规则,第 3 列表示目标词应该标注的词性,第 4 列表示目标词标注成某一词性的概率。

3)对第 2)步所得的文件里面的数据进行处理,将目标词的规则读入内存。

4)根据顺序对规则进行解析、匹配。

由于本文所用的是目标词前后不同位置的词语信息作为规则,所以匹配的时候存在一个优先级问题。本文匹配规则的顺序是按照式(7)概率的大小来进行的。即将概率较大的触发对赋予较高的优先级,匹配的时候优先考虑。

5)最后根据匹配的结果确定兼类词的词性标注结果。

4　实验与结果分析

实验语料采用的是 1998 年 1 月的《人民日报》语料和从《语料库在线》上获得的语料,并对语料进行人工校对后作为实验用的标准语料,分别进行测试。《人民日报》语料库是比较成熟的语料库,相对而言质量比较好,《语料库在线》由于没有经过专家的校对,质量稍差。对于 CRF 模型和 ME 型,本文把上下文窗口的有效范围控制在(-4,4)。

条件随机场使用的是 CRF++(其中 ++ 表示版本信息)工具包。最大熵模型采用的是张乐博士开发的最大熵工具包,先把训练语料和测试语料处理成工具包要求的格式,然后对处理好的语料进行训练和测试。

实验分为两部分:规则挖掘实验和词性标注实验。

4.1 规则挖掘实验

表 1 给出了部分规则统计的实例。其中:*A* 代表规则单元词,*B* 代表目标词单元,*MI* 表示互信息的大小。

从表 1 中给出的词对来看,互信息统计算法得出的规则很好地体现了语言学上的知识,由规则词就能够确定目标词的词性,统计得出的规则可以作为本文目标词的词类划分依据。

表 1 部分规则实例

A	*B*	*MI*	*A*	*B*	*MI*
努力	工作/v	1.000	大力	发展/v	1.000
的	工作/vn	0.992	进一步	提高/v	0.969
把	建设/v	1.000	深入	学习/v	1.000

本文所选目标词的规则数统计如表 2 所示。表 2 分别给出了目标词根据互信息统计算法和人工观察的规则条数,分别在两种语料库上进行实验,R 表示《人民日报》语料,Y 表示《语料库在线》语料。

表 2 规则数统计

兼类词	互信息统计算法		人工统计	
	Y	R	Y	R
发展(v/vn)	73	64	83	71
服务(v/vn)	29	19	35	23
工作(v/vn)	111	98	123	114
建设(v/vn)	83	68	91	79
生产(v/vn)	23	15	27	21
使用(v/vn)	15	10	19	16
提高(v/vn)	37	28	42	33
学习(v/vn)	25	12	29	17
研究(v/vn)	26	20	31	23
影响(v/vn)	37	14	43	21

从表 2 可以看出,本文的方法对于词性标注规则的获取具有比较好的效果,基本上能全面覆盖每个词的所有词性标注规则,大幅度减少了人工编写规则的工作量。每个兼类词的规则条数都不一样。这是因为本文训练语料中所包含目标词的句子条数不一样,而且汉语词语的用法非常灵活,每个词的用法和搭配都不一样,有些词语的固定搭配比较多,这类词更容易找到兼类词的标注规则,有些词的固定搭配比较少,和很多词都可以共现,此时就很难找到词性标注规则。

4.2 词性标注实验

3 种模型开放测试兼类词词性标注准确率对比的结果如表 3 所示,分别在两种语料库上进行实验,其中:R 表示《人民日报》语料库,Y 表示《语料库在线》语料库,分别取其中一部分作为训练语料,另外一部分作为测试语料。

从表 3 的结果可以看出,在用于基于最大熵方法对兼类词进行识别,选取相同的特征模板时,正确率比较高,总体上优于 HMM 和 CRF 模型,并且性能稳定。

HMM 和 CRF 模型虽然也取得较好的标注效果,但由于其预测信息的不足,对词性标注,特别是对兼类词和未登录词的词性标注准确率影响很大。而最大熵模型使用特征的形式,有效地利用了上下文信息,在一定的约束条件下可以得到与训练数据一致的概率分布;即使是未登录词,由于其丰富的上下文信息,对它的词性标注也起到了很好的预测作用。实验结果证明最大熵方法取得了比上述模型较好的标注效果。

表 3 统计模型标注准确率对比

兼类词	CRF		ME		HMM	
	Y	R	Y	R	Y	R
发展	0.823	0.859	0.845	0.889	0.704	0.728
服务	0.799	0.820	0.813	0.878	0.723	0.745
工作	0.843	0.889	0.852	0.913	0.855	0.902
建设	0.855	0.912	0.912	0.913	0.831	0.879
生产	0.773	0.742	0.821	0.863	0.721	0.763
使用	0.814	0.861	0.833	0.907	0.811	0.869
提高	0.879	0.928	0.881	0.948	0.879	0.932
学习	0.864	0.855	0.873	0.853	0.834	0.860
研究	0.812	0.816	0.813	0.856	0.714	0.785
影响	0.892	0.891	0.901	0.881	0.729	0.737

总体来说,在《人民日报》语料库上进行词性标注实验时,标注准确率比较高,这是由于该语料库人工标注的准确率相对而言比较高,更能体现汉语语言的概率分布特征,这说明语料库的质量也是影响统计模型汉语词性标注时的准确率因素之一。

统计模型加入规则之后的标注准确率如表 4 所示,分别在两种语料库上进行实验,其中:R 表示《人民日报》语料库,Y 表示《语料库在线》语料库。

表 4 统计模型加规则后标注准确率对比

兼类词	CRF + 规则		ME + 规则		HMM + 规则	
	Y	R	Y	R	Y	R
发展	0.903	0.921	0.891	0.911	0.864	0.892
服务	0.879	0.892	0.879	0.902	0.879	0.928
工作	0.899	0.933	0.913	0.942	0.923	0.952
建设	0.913	0.954	0.921	0.933	0.917	0.954
生产	0.854	0.863	0.873	0.895	0.899	0.902
使用	0.879	0.914	0.942	0.951	0.963	0.971
提高	0.923	0.967	0.951	0.964	0.931	0.962
学习	0.883	0.904	0.875	0.892	0.922	0.939
研究	0.824	0.857	0.954	0.871	0.913	0.921
影响	0.897	0.921	0.892	0.915	0.932	0.944

通过表 4 和表 3 的对比,本研究发现加入规则之后,词性标注的准确率都得到了一定的提升,基本上都达到了 90% 以上。基于统计的方法是单纯地依靠概率来推断一个词的词性,准确率与训练语料的质量和训练语料的规模关系很大。有很多时候仅仅依靠统计的概率很难正确地给定一个词的词性标注。规则更多体现的是语言学上的知识,不依赖于训练语料,规则具有更高的准确性,可以作为统计词性标注模型的一个补充和优化的过程。实验结果表明,基于互信息的规则发现方法是有效的,对于词性标注结果的优化具有很好的效果。

5 结语

本文主要使用了基于条件随机场、最大熵、隐马尔可夫 3 种统计方法对常用的兼类词进行识别研究,并针对不同的方法分别考虑了兼类词本身的特点以及在上下文中的词语以及词性对其产生的影响。由于传统的统计方法对兼类词标注的

准确率不高,本文将传统的统计模型结合统计规则对兼类词进行词性标注,并介绍了规则的形式化描述及基于规则的兼类词识别算法,实验结果表明了改进的互信息算法在规则的获取上的有效性。下一步的工作是针对更多的兼类词尝试用规则的方法进行识别,完善规则库,并且尝试用聚类的方法对兼类词的识别进行研究。

参考文献:

[1] BRILL E. A corpus-based approach to language learning [D]. Philadelphia: University of Pennsylvania, 1993.

[2] HAMMERTON J, OSBORNE M, ARMSTRONG S, *et al*. Introduction to special issue on machine learning approaches to shallow parsing [J]. Journal of Machine Learning Research, 2002, 13(2): 551 – 558.

[3] BRILL E. Unsupervised learning of disambiguation rules for part-of-speech [C]// Proceedings of the Third Workshop on Very Large Corpora. Piscataway: IEEE Press, 1995: 1 – 13.

[4] SCHMID H. Probabilistic part-of-speech using decision tree [C]// Proceedings of the 1994 International Conference on New Methods in Language Processing. Piscataway: IEEE Press, 1994: 44 – 49.

[5] YUAN C. Improved hidden Markov model for speech recognition and POS tagging [J]. Journal of Central South University, 2012, 19(2): 511 – 516.

[6] RATNAPARKHI A. A maximum entropy model for part-of-speech tagging [C]// Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1996: 132 – 142.

[7] LIU T, LEI L, CHEN L. A parallel training research of Chinese part-of-speech tagging CRF model based on MapReduce [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(1): 147 – 152. (刘滔, 雷霖, 陈荦, 等. 基于 MapReduce 的中文词性标注 CRF 模型并行化训练研究[J]. 北京大学学报: 自然科学版, 2013, 49(1): 147 – 152.)

[8] EMBAL A, SAHA S. Simulated annealing based classifier ensemble techniques: application to part of speech tagging [J]. Information Fusion, 2013, 14(3): 288 – 300.

[9] ZHAO Y, WANG X, LIU B, *et al*. Fusion of clustering trigger-pair features for POS tagging based on maximum entropy model [J]. Journal of Computer Research and Development, 2006, 43(2): 268 – 274. (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).

[10] BRANTS T. TnT: a statistical part-of-speech tagger [C]// Proceedings of the 6th Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2000: 224 – 231.

[11] COHEN W, CARVALHO V. Stacked sequential learning [C]// IJCAI'05: Proceedings of the 19th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2005: 671 – 676.

[12] ZHAN Y, CLARK S. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model [C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2010: 843 – 852.

[13] AULI M, LOPEZ A. Training a log-linear parser with loss functions via softmax-margin [C]// EMNLP'11: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 333 – 343.

(上接第 2196 页)

[4] W3CSchool. OWL introduction [EB/OL]. [2013-09-18]. http://www.w3school.com.cn/rdf/rdf_owl.asp. (W3CSchool. OWL 简介 [EB/OL]. [2013-09-18]. http://www.w3school.com.cn/rdf/rdf_owl.asp.)

[5] Movieontology.org. MO — the movie ontology [EB/OL]. [2013-09-18]. http://www.movieontology.org/.

[6] USCHOLD M, KING M, MORALEE S, *et al*. The enterprise ontology [J]. The Knowledge Engineering Review, 1998, 13(1): 31 – 89.

[7] University of Toronto, Faculty of Applied Science and Engineering. TOVE ontologies [EB/OL]. [2013-09-18]. http://www.ie.utoronto.ca/EIL/tove/toveont.html.

[8] NOY N F, McGUINNESS D L. Ontology development 101: a guide to creating your first ontology: knowledge systems laboratory, SMI-2001-0880 [R]. Stanford: Stanford University, 2001.

[9] Editorial Committee of China Library Classification. China library classification [M]. 4th ed. Beijing: Beijing Library Press, 1999: 197 – 199. (中国图书馆分类法编辑委员会. 中国图书馆分类法 [M]. 4 版. 北京: 北京图书馆出版社, 1999: 197 – 199.)

[10] XIA Z, CHEN Z. Unabridged dictionary: art [M]. 6th ed. Shanghai: Shanghai Lexicographical Publishing House, 2010. (夏征农, 陈至立. 辞海: 艺术分册[M]. 6 版. 上海: 上海辞书出版社, 2010.)

[11] Editorial Department in Encyclopedia of China Publishing House. Encyclopedia of China: movie [M]. 2nd ed. Beijing: Encyclopedia of China Publishing House, 2004. (中国大百科全书出版社编辑部. 中国大百科全书: 电影[M]. 2 版. 北京: 中国大百科全书出版社, 2004.)

[12] Douban. Douban movie [EB/OL]. [2014-04-09]. http://movie.douban.com/. (豆瓣. 豆瓣电影[EB/OL]. [2014-04-09]. http://movie.douban.com/.)

[13] IMDb Chinese Website. Movie database [EB/OL]. [2014-04-09]. http://www.imdb.cn/. (IMDb 中文网. 电影资料库[EB/OL]. [2014-04-09]. http://www.imdb.cn/.)

[14] Stanford Center for Biomedical Informatics Research. Protégé [EB/OL]. [2014-04-09]. http://protege.stanford.edu/.

[15] USCHOLD M, GRUNINGER M. Ontologies: principles, methods and applications [J]. The Knowledge Engineering Review, 1996, 11(2): 93 – 136.

[16] GAO Y, CAO C, SUI Y. Musical domain-specific ontology building and analysis [J]. Computer Science, 2004, 31(1): 103 – 107. (高颖, 曹存根, 眭跃飞. 音乐领域本体的建立和分析[J]. 计算机科学, 2004, 31(1): 103 – 107.)

[17] XIE N, WANG W. Ontology and acquiring of agriculture knowledge [J]. Agriculture Network Information, 2007(8): 12 – 16. (谢能付, 王文生. 农业知识本体构建方法[J]. 农业网络信息, 2007(8): 12 – 16.)