半监督学习的老挝语词性标注方法研究

杨 蓓 周兰江 余正涛 刘丽佳

(昆明理工大学信息工程与自动化学院 昆明 650500)

(昆明理工大学智能信息处理重点实验室 昆明 650500)

摘 要 针对老挝语语料资源极少而无法直接利用有监督学习的方法实现老挝语词法分析的问题,提出了基于半监督学习的老挝语词性标注方法。首先利用仅有的少量标注词典和未标注语料资源,采用简单概率模型建模,获取较为完整的标注词典;其次利用整数规划获取大量自动标注的语料;最后在训练语料充足的情况下,利用二阶隐马尔科夫模型建模,实现高质量的老挝语词性标注。提出的方法在老挝语词性标注方面取得了较好的效果,其准确率达到89.8%。

关键词 半监督学习,二阶隐马尔科夫模型,老挝语词性标注,概率模型,整数规划

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.019

Research on Semi-supervised Learning Based Approach for Lao Part of Speech Tagging

YANG Bei ZHOU Lan-jiang YU Zheng-tao LIU Li-jia

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

(The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming 650500, China)

Abstract Aiming at the problem of very few corpora resources, a semi-supervised learning based approach for Lao part of speech Tagging was presented. Firstly, a simple probability model is used to obtain a complete dictionary with a small amount of tagged dictionary and untagged corpus, then much more automatically tagged corpus with integer programming are obtained. Finally, a second-order Markov model with sufficient corpus resources is trained to realize a high quality Lao part of speech tagging. This method achieves a good result in Lao part of speech tagging, and its accuracy is up to 89.8%.

Keywords Semi-supervised learning, Second-order hidden markov model, Lao part of speech tagging, Probability model, Integer programming

1 引言

词性标注指根据句子上下文中的信息给句中的每个词一个正确的词性标记,这些标记都有特定的语言学意义。词性标注工作作为老挝语命名实体识别、依存句法分析、词义消歧、语义角色标注等研究工作的重要基础,已被应用于如文本索引、文本分类、语料库加工等众多领域,因此研究词性标注的方法具有重要意义。老挝语是一种孤立型语言,这类语言的特点在于其一般不是通过词的内部形态变化(又称作屈折变化)来表达语法的作用,而是通过独立的虚词和固定的词序来表达语法意义,而且一般而言,分析语缺乏多数的格变化,但是老挝语有丰富的词缀来改变词性或者词义。在词法方面,老挝的词同汉语类似,也分为名词、代词、动词、形容词、副词、介词、连词、叹词等。这些词汇有老挝语的原词,还有相当一部分的借词,其中有巴利语借词、梵语借词、高棉语借词、泰语借词、汉语借词和法语借词。由于历史和地理等原因,老挝语受高棉语(柬埔寨语)的影响最大。作为低资源语言的老挝语,国内外对其研究起步较晚,目前没有针对老挝语词性标注的研究;对于英语、汉语等语言的词性标注方法研究已经相对成熟;而对其他低资源语言,如对卢旺达语、越南语等的词性标注,国内外近年来也有一些研究。洪铭材[1]提出采用条件随机场,针对兼类词和未登录词添加了新的统计特征,标注的准确率达到98.56%;Dan Garrette[2]提出利用隐马尔科夫对标注词典建模,定义一个简单的隐马尔科夫发射初始值,在给定标注的基础上,估计生语料的标注结果,在英语和意大利语上都得到了很好的标注效果;王丽杰[3]利用SVMTool对中文进行词性标注,在模型训练时加入部首特征和词重叠特征,最终词性标注的准确率比基线系统提高了2.0%;Bernard Merialdo[4]采用概率模型来标注英文,该文献的创新之处在于利用未标注的语料来训练模型,通过在大量人工标注语料

到稿日期:2015-08-01 返修日期:2015-09-08 本文受面向汉语-泰语跨语言新闻事件检索方法研究(61462054)资助。

杨 蓓(1989-),女,硕士生,主要研究方向为自然语言处理与嵌入式系统研究,E-mail:likeseayb@163.com;周兰江(1964-),男,副教授,主要研究方向为自然语言处理与嵌入式系统研究,E-mail:915090822@qq.com;余正涛(1964-),男,教授,博士生导师,主要研究方向为自然语言处理等,E-mail:ztyu@hotmail.com;刘丽佳(1988-),女,硕士生,主要研究方向为自然语言处理与信息抽取,E-mail:839856859@qq.com。

的基础上计算相关标注频率来估计模型参数，取得了较好的标注效果。

本文借鉴成熟的词性标注方法，利用人工标注完成的老挝语词典，选择概率模型对词缀和词性建模，以获取较为完整的词典，最后融入老挝语的词性特征来实现老挝语的词性标注。

2 较完整老挝语词典的获取

2.1 老挝语词性标注集

老挝文字是由元音、辅音、尾辅音和声调符号组成的拼音文字。老挝语与汉语语句中主要成分的排列顺序是一样的，即主、谓、宾3大成分的排列顺序是一样的，都是主＋谓＋宾的顺序。词的类别和汉语类似，主要有名词、代词、动词、形容词、量词等。老挝语中名词主语句中的主语都可以由名词、数词、代词或名词性短语充当；宾语都是由名词构成；充当句子中谓语成分的词或词组可以是动词、形容词、动词性短语或形容词性短语以及名词；谓语成分和汉语一样都是位于主语后的。老挝语人名的前面一般加上称谓。

目前没有老挝语的词性标注体系，本文根据老挝语的语言特点，参照哈尔滨工业大学语言技术平台LTP的中文词性标注体系，将所有单词分为30种词性，部分词性如表1所列。

表1 老挝语词性标注集

序号	标注	描述	举例
1	a	adjective	ງາມ
2	c	conjunction	ແລະ
3	d	adverb	ຫຼາຍ
4	e	exclamation	ອະນິຈາ
5	i	idiom	ນິຍົມ
6	r	abbreviation	ຈີນພັກຮ່ວມມູນິດ
7	m	numeral	ໜຶ່ງ, ຄັ້ງທຳອິດ
8	n	general noun	ເຊົ້າຈີ
9	dn	direction noun	ໄວ້
10	pn	Person name	ບົວສອນບຸບຜາວັນ
11	ln	location name	ຈີນ
12	on	organization name	ອົງການການຄ້າໂລກ
13	q	quantifier	ເດືອນ
14	p	preposition	ໃນ
15	v	verb	ການດຳເນີນງານ
16	pr	pronoun	ພວກເຮົາ
17	om	onomatopoeia	ຝົນຕົກແຮງ
18	au	auxiliary	ຂອງ, ດີນ
19	pu	punctuation	
20	X	non-lexeme	

2.2 老挝语词汇形态特征

老挝语的词汇有一定的词汇形态规律，通过上下文的形态信息可以对指定词进行合理分析。例如：老挝语有些量词还有名物化的作用，即在某些形容词的前面加上一定的量词其就变成名词了。如在形容词“熟”前面加上量词“个”变成“熟的”，在“大”前面加上“个”变成“大的、大个的”，由此可推测，老挝语中的许多量词都是由名词发展而来的；在老挝语现在进行时句子中会在动词前加上ພວມ、ກຳລັງ，句子末尾会加上ຢູ່[2]。

以老挝语动词的时态变化形态规律为例进行总结，如表2所列。

表2 老挝语动词形态

动词时态	动词形态	例句
现在进行时	动词前加ພວມ、ກຳລັງ，句子末尾加ຢູ່。 ພວມ+V.....ຢູ່或者ກຳລັງ+V.....ຢູ່	ຂ້ອຍ ກຳລັງ(标志词) ກິນເຂົ້າ (v) ຢູ່ 我 正在 吃饭。
过去时	动词前加ໄດ້, ເຄີຍ... ໄດ້+V 或 ເຄີຍ+V	ລາວໄດ້ (标志词) ມາຫາ(V) ເຈົ້າສອງເທື່ອແລ້ວ. 他已经来找过你两次了。
	放在动词之后或句尾的词：ແລ້ວ V+ ແລ້ວ	ຂ້ອຍກິນ(V)ແລ້ວ(标志词). 我 吃 过了。
将来时	动词之前加ຈະ或ຈະ ໃຫ້... V+ຈະ或 V+ຈະ或 V+ ໃຫ້	ເຈົ້າຈະ(标志词) ໄປ(V)ໃສ ？ 你要去哪里？
	放在句末ແລ້ວ..	ເພິ່ນຈະມາ(V) ແລ້ວ(标志词). 他就要来了。

2.3 根据老挝语词汇特征获取较完整词典

由于标注词典的数量较少，这样直接训练模型来标注生语料，将导致不完整的标注词典和未登录词使标注准确率很低。为了获取较为完整的标注词典，减少未登录词的出现，首先从标注词典中抽取常用的词缀，组成包含50个词缀的词缀集[5]，使用词典中的词和标注对，训练一个简单的概率模型 $p(tag|affix)$，然后将这个模型应用到未登录词的词性标注上，从而扩大词典量，获取较完整的标注词典。

3 利用整数规划生成标注语料

整数规划(Integer Programming，IP)是在有限个可供选择的方案中寻找满足一定约束的最好方案。将整数规划应用到词性标注问题上，就是通过条件约束最终选择符合要求的标注序列。由于生成的标注语料将用于训练二阶隐马尔科夫(Second-order Hidden Markov Model)模型的语料，在训练模型的过程中，语法模型越小，训练时间就越短[6]，因此选择可以解释观察序列的最小语法模型作为标注结果。

在词典中，每个老挝语词对应它所有的词性标记及该标记的频率，通过该标签词典对给定的老挝语词序列进行整数规划[7]，利用CPLEX软件来解决这个过程中的标记序列的大量组合问题。对老挝语ເຈົ້າ ຊິ ໄປ ໃສ？(你要去哪里？)的整数规划过程如图1所示。

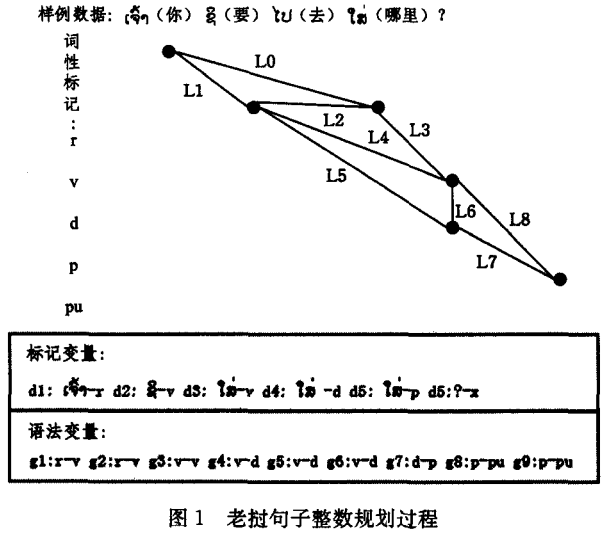

图1 老挝句子整数规划过程

整数规划算法：如图 1 所示，创建一个网络结构，该结构的每个节点代表所表示词在词典中对应的一种词性，对每一条边都赋予一定的权重，进入任意一个节点的边权重和等于离开这个节点的边权重和，例如：$L0=L2+L3$。同时对每个标记变量、语法变量赋值。每个边变量都满足语法变量与标记变量的约束。例如只有边变量所受的语法变量满足"ก่าลัง"才能给该边赋"ก่าลัง"的权值。目标函数：$\sum_{i=1\cdots10} gi$ 取得最小值，循环执行以下两步操作：

(1)每次只选择一条边，该边满足从左到右(而不是其它)，每个节点的输入权值等于输出权值；

(2)选择的边同时满足标记变量与语法变量约束。

循环执行以上操作直到 $\sum_{i=1\cdots10} gi$ 取得最小值时结束，生成该老挝语词序列对应的词性标记序列集合，在该集合中随机选取一种作为最终词性标注结果。IP 模型获得的标注可能是不准确的，直接采用其结果应该会对 HMM 模型带来不利影响，因此在标记集合中删去有问题(歧义、冗余、噪音)的语料，实现老挝语语料词性自动标注。

4 基于二阶隐马尔科夫模型的词性标注研究

4.1 老挝语 N 元语法模型

N 元语法模型认为自然语言知识可以用连续的符号序列(如字序列、词序列、词性标记序列、语音波形序列等)的概率来表示[8,9]。通过 N 元语法模型表示老挝语的词性标记序列，其基本思想是：假设老挝语的文本的产生服从马尔科夫链，且第 n 个语法单位只和紧挨着它的前面很少的 $n-1$ 个词有关，则可以根据从在之前工作中得到的词性自动标注语料训练提取的参数，计算词串可能对应的词性标记串的概率，实现对老挝语文本词性的概率表示。

对于老挝词性标记串 $W=W_1,W_2,W_3,\cdots,W_n$ 来说，可以认为老挝词性 W_i 的出现与整个上文的老挝语词性 $W_1,W_2,\cdots,W_{n-1}$ 相关。那么老挝语词性串 W 出现的概率可以通过以下方法得出：

$$\begin{aligned}P(w)&=P(w_1w_2\cdots w_n)\\&=P(w_1)P(w_2|w_1)\cdots P(w_n|w_1\cdots w_{n-1})\\&=\sum_{i=1}^{n}P(w_i|w_{i-n+1}\cdots w_{i-2}w_{i-1})\end{aligned}\tag{1}$$

其中，$P(w_n|w_1\cdots w_{n-1})$ 表示在已知先前词性序列的条件下选取老挝词性 W_n 的概率。这样即使在 i 不是很大的情况下也很难计算概率 P，因为这需要计算太多统计信息而使得模型过于复杂。可以适当地忽略一些对当前老挝词词性结果不产生影响或影响很小的历史信息，从而来简化模型。假设老挝词性 W_i 只与上文的前 n 个词性(一般为当前词在老挝句子位置以前的老挝词词性)有关，则式(1)变为：

$$P(w)=P(w_1w_2\cdots w_n)=\sum_{i=1}^{n}P(w_i|w_{i-n+1}\cdots w_{i-2}w_{i-1})\tag{2}$$

式(2)就是定义的老挝语词性 N 元语法模型，其中，当 n 取值为 1,2,3 时，分别叫做一元词性老挝语语法模型(Laou-nigram)、二元词性老挝语语法模型(Lao-bigram)和三元词性老挝语语法模型(Lao-trigram)。

4.2 二阶隐马尔科夫模型

传统的隐马尔科夫模型在计算词性标注的最佳序列时，仅仅考虑了上文词性对当前词词性的影响，这就使得一些重要的上下文信息发生了丢失，在老挝语词性方面，有些词的词性与下文词性有关。针对隐马尔科夫模型在词性标注上的缺陷，本文采用了一种基于上下文的二阶隐马尔科夫模型[10,11]。

由于传统隐马尔科夫模型中参数的定义仅与上文有关，因此要做到与上下文有关，首先需要重新定义模型参数，故在这里定义基于上下文的二阶隐马尔科夫模型中各参数为：$\theta=(N,M,\pi,A,B)$。

1)N 表示标注语料中所用的标记集词性的个数；

2)M 表示所用词汇集词汇的个数；

3)π_i 表示词性 t_i 作为句首出现的概率；

4)词性概率转移矩阵 $A=\{a_{ijk}\}$，其中：

$$a_{ijk}=P(s_m=t_j|s_{m-1}=t_i,s_{m+1}=t_k),\quad 3\leqslant m\leqslant M-1,1\leqslant i,j,k\leqslant N\tag{3}$$

其中，t_j 的词性转移概率不仅依赖于上文 c_{m-1} 的词性 t_i，而且也与下文 c_{m+1} 的词性 t_k 有关，这样就可以将老挝语词性的上下文信息都考虑在内；

5)词汇概率矩阵 $B=\{b_{ij}(w_k)\}$，其中：

$$b_{ij}(w_k)=P\{o_m=w_k|s_m=t_i,s_{m+1}=t_j\},\quad 1\leqslant m,k\leqslant M,1\leqslant i,j\leqslant N\tag{4}$$

其中，$b_{ij}(w_k)$ 表示在 s_m 的状态为 t_i 且 s_{m+1} 在状态 t_j 的条件下，o_m 输出为 w_k 的概率。

由于对传统隐马尔科夫模型参数做了上述的调整，因此也需要对 Viterbi 算法进行改进，以更好地用于改进后的模型。

初始化：

$$\delta_1(i,j)=\pi_ib_{ij}(o_1),1\leqslant i,j\leqslant N\tag{5}$$

$$j_1(i,j)=0,1\leqslant i,j\leqslant N\tag{6}$$

递归：

$$\delta_m(j,k)=\max_{1\leqslant i\leqslant N}[\delta_{m-1}(i,j)a_{ijk}]b_{jk}(o_m),\quad 2\leqslant m\leqslant M-1,1\leqslant j,k\leqslant N\tag{7}$$

$$j_m(j,k)=\arg\max_{1\leqslant i\leqslant N}[\delta_{m-1}(i,j)a_{ijk}],\quad 2\leqslant m\leqslant M-1,\quad 1\leqslant j,k\leqslant N\tag{8}$$

最后：

$$p^*=\max_{1\leqslant j,k\leqslant N}[\delta_{M-1}(j,k)]c_k(o_M)\tag{9}$$

$$q_M^*=\arg\max_{1\leqslant j,k\leqslant N}[\delta_{M-1}(j,k)]c_k(o_M)\tag{10}$$

由以上修改 Viterbi 算法的改进后的二阶隐马尔科夫模型作为老挝语词性标注训练模型，其考虑了老挝语词性上文信息与下文信息，提高了老挝语词性标注的效果。

5 实验与分析

影响词性标注系统标注正确率的因素主要有[12]：1)训练语料库规模的大小和语料内容涉及的领域；2)词性标记集的大小，词性标记集的划分越细，标注的正确率就会降低；3)训练语料库是人工标注的，而且训练语料的规模也比较大，因此难免会出现错误的标注，这也是影响标注正确率的一个因素。

在实验中，采用了文中定义的 30 种词类组成的标注集，标注词典中共有约 30000 词条，语料是从中国百科和老挝语相关网站上获取的，包含约 60 万词，内容涉及政治、经济、地理、体育等题材。在词典数量不变的条件下，分别对 10 万，20 万，…，60 万词的语料进行开放和封闭测试。

5.1 不同规模的训练语料实验

在进行封闭测试时，首先分别以 10 万，20 万，…，50 万词的语料进行训练，建立相应的二阶隐马尔科夫模型，然后采取改进的 Viterbi 算法对用来训练的所有语料重新进行词性标注，求出每一个句子的最佳词性标注序列，即完成了封闭测试。

在进行开放测试时，先从 60 万的语料库中抽出 10%的句子，这些句子不参与训练，用于后面的测试。同样地，分别以 10 万，20 万，…，50 万词的语料进行训练，建立相应的语料词典和二阶隐马尔科夫模型参数，然后采取改进的 Viterbi 算法对测试集进行词性标注，从而完成了开放测试。

封闭测试和开放测试的准确率与训练语料规模的关系如图 2 所示。

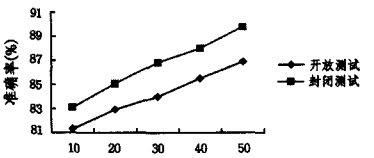

图 2 封闭测试和开放测试时准确率与语料规模的关系

5.2 词性标注方法的比较

标注语料资源有限，在选择词和词缀特征方面有一定的局限性，因此采用 SVMTool 和概率模型会影响标注结果，而隐马尔科夫模型克服了这些局限性。二阶隐马尔科夫模型考虑了更多的上下文信息，提高了标注的准确性，将利用本文的方法与利用 SVMTool、概率模型和一阶隐马尔科夫建立词性标注模型进行了比较，实验结果如表 3 所列。

表 3 本文方法与几种方法的比较结果(%)

方法	准确率 P	召回率 R	F 值
SVMTool	76.2	72.9	74.5
概率模型	79.5	78.3	78.9
一阶隐马尔科夫模型建模	83.9	77.7	80.7
本文方法	89.8	81.2	85.3

实验结果显示，使用本文方法在老挝语词性标注的过程中准确率比较高，达到 89%以上，说明采用概率模型扩大词典数量、增大标注语料的数量，同时考虑上文词性对当前词词性的影响为提高标注准确率提供了有力保障。

结束语 本文在语料有限的条件下，首先利用少量的标注词典和未标注语料，采用概率建模的方法，获取数量较多的词典，然后采用整数规划获取大量标注语料，经过人工去除有噪音、歧义的语料，产生较为规范的标注语料，最后采用二阶隐马尔科夫模型，利用改进后的 Viterbi 算法完成了对老挝语的自动词性标注。分别以 10 万，20 万，…，60 万词级的语料库和 30000 词条的标注词典作为训练语料，进行了开放和封闭测试，将本文的方法和几种词性标注方法进行了比较，由结果可知扩大训练语料后标注准确率会提高；同时与几种方法相比，利用本文方法进行词性标注时标注准确率也有所提高，达到 89.8%。但是这与应用系统的要求还有一定的差距。为了提高标注系统的正确率，需要进一步研究如何将老挝语的切分与隐马尔科夫模型结合起来进行词性消歧。

参 考 文 献

[1] Hong Ming-cai, Zhang Kuo, Tang Jie, et al. A Chinese Part-of-Speech Tagging method based on conditional random fields (CRFs)[J]. Computer Science, 2006, 33(10): 148-155(in Chinese)
洪铭材，张阔，唐杰，等. 基于条件随机场(CRFs)的中文词性标注方法[J]. 计算机科学，2006，33(10)：148-155

[2] Dan G, Baldridge J. Type-Supervised Hidden Markov Models for Part-of-Speech Tagging with Incomplete Tag Dictionaries[C]// Proceedings of the Association for Computational Linguistics (ACL). 2012: 821-831

[3] Wang Li-jie, Che Wan-xiang, Liu Ting. Chinese Part-of-Speech Tagging Based on SVMTool[J]. Journal of Chinese Information Processing, 2009, 23(4): 16-21(in Chinese)
王丽杰，车万翔，刘挺. 基于 SVMTool 的中文词性标注[J]. 中文信息学报，2009，23(4)：16-21

[4] Merialdo B. Tagging english text with a probabilistic model[J]. Computational Linguistics, 2002, 20(2): 155-171

[5] Garrette, Baldridge J. Learning a Part-of-Speech Tagger from Two Hours of Annotation[C]// Proceedings of the Association for Computational Linguistics(ACL). 2013: 138-147

[6] Toutanova K, Johnson M. A Bayesian LDA-based model for semi-supervised part-of-speech tagging [C] // Proceedings of The Annual Conference on Neural Information Processing Systems(NIPS). 2008: 1521-1528

[7] Ravi S, Knight K. Minimized Models for Unsupervised Part-of-Speech Tagging[C]// Proceedings of the Association for Computational Linguistics(ACL). 2009

[8] Liang Yi-min, Huang De-gen. Full second-order Hidden Markov model based Part-of-Speech Tagging[J]. Computer Engineering, 2005, 31(10): 177-180(in Chinese)
梁以敏，黄德根. 基于完全二阶隐马尔可夫模型的词性标注[J]. 计算机工程，2005，31(10)：177-180

[9] Liu Jie-bin, Song Mao-qiang, Zhao Fang, et al. Context based second-order Hidden Markov model[J]. Computer Engineering, 2010, 36(10): 231-235(in Chinese)
刘洁彬，宋茂强，赵方等. 基于上下文的二阶隐马尔可夫模型[J]. 计算机工程，2010，36(10)：231-235

[10] Feng Yue-jiao, He Xing-shi. Theory and Implementation of second-order Hidden Markov model[J]. Value Engineering, 2009 (12): 103-105(in Chinese)
丰月姣，贺兴时. 二阶隐马尔科夫模型的原理与实现[J]. 价值工程，2009(12)：103-105

[11] Thede S M, Harper M P. A second-order Hidden Markov Model forpart-of-speech tagging[C]// Proceedings of the Association for Computational Linguistics(ACL). 1999: 20-26

[12] Yang Hong, Wng Sigerileng. HMM based Automatic Mongolia Part-of-Speech Tagging[J]. Journal of Inner Mongolia Normal University(Natural Science Edition), 2010, 39(2): 206-209(in Chinese)
艳红，王斯日古楞. 基于 HMM 的蒙古文自动词性标注研究[J]. 内蒙古师范大学学报(自然科学汉文版)，2010，39(2)：206-209