文章编号：

基于层叠条件随机场的高棉语分词及词性标注方法*

潘华山[1,2]，严馨[1,2]，余正涛[1,2]，郭剑毅[1,2]，李王[1]，庄智[1]

（1.昆明理工大学信息工程与自动化学院，云南 昆明 650500；

2.昆明理工大学智能信息处理重点实验室，云南 昆明 650500）

摘要：针对高棉语分词及词性标注问题，提出一种基于层叠条件随机场模型的自动分词及词性标注方法。该方法由三层条件随机场模型构成：第一层是分词模型，该模型以KCC为粒度，结合上下文信息与高棉语的构词特点构建特征模板，实现对高棉语句子的自动分词；第二层是分词结果修正模型，该模型以词语为粒度，结合上下文信息与高棉语中命名实体的构成特点构建特征模板，实现对上层模型的分词结果中的命名实体切分错误进行修正；第三层是词性标注模型，该模型以词语为粒度，结合上下文信息与高棉语丰富的词缀信息构建特征模板，实现对高棉语句子中的词语进行自动标注词性。基于该模型进行开放测试实验，最终准确率为95.44%，结果表明提出方法能有效解决高棉语的分词和词性标注问题。

关键词：高棉语；层叠条件随机场；分词；词性标注

中图分类号：TP391　　**文献标识码：**A

A Khmer Word Segmentation and Part-Of-Speech Tagging Method Based on Cascaded Conditional Random Fields

PAN Huashan[1,2], YAN Xin[1,2], YU Zhengtao[1,2], GUO Jianyi[1,2], LI Wang[1], ZHUANG Zhi[1]

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

Abstract: For the Khmer word segmentation and Part-Of-Speech(POS) tagging problem, we proposed a Khmer automatic word segmentation and POS tagging method based on Cascaded Conditional Random Fields(CCRFs) model. The approach consists of three layers of Conditional Random Fields(CRFs) models: the first layer is the word segmentation model in KCC granularity which combined with the context of Khmer text to build the feature template to realize the Khmer automatic word segmentation of sentences; the second layer is the word segmentation result correcting model in word granularity, which combined with the Khmer context and the forming characteristic of Khmer named entities to build the feature template to correct the improper segmentation of named entities from the upper word segmentation model; the third layer is the POS tagging model in word granularity, which combined with the Khmer context and rich affixes information to build the feature template, and achieved the Khmer POS tagging. We made a group of contrast test based on the model on an open corpus, and obtained a final accuracy rate of 95.44 percent, the result shows that the proposed method can effectively solve the Khmer word segmentation and POS tagging problems.

Keywords：Khmer;cascaded conditional random fields;word segmentation;part-of-speech tagging

1 引言

柬埔寨语又称高棉语，属南亚语系孟高棉语族高棉语支语言，是柬埔寨现在的官方语言。高棉语是在古高棉语的基础上演变和发展而来的，由于在历史上与印度宗教之间的密切关

*收稿日期：2013-08-30 定稿日期：2013-10-09

基金项目：汉越双语语料库建设及词对齐方法研究（No.6126041）

作者简介：潘华山（1987——），男，在读研究生，研究方向为自然语言处理、数据挖掘；严馨（1969——），女，副教授，研究方向为自然语言处理、数据挖掘；余正涛（1971——），男，教授，研究方向为自然语言处理、问答系统、信息检索；郭剑毅（1964——），女，教授，研究方向为信息抽取；李王（1989——），男，本科；庄智（1990——），男，本科。

系，高棉语吸收了许多巴梵语词汇，同时又与周边国家语言发生接触，如汉语、泰语、越南语、老挝语等。近代由于法国殖民统治和现代科技的发展，也吸收了许多英法词汇。因此高棉语的构词方式比较多样化。其主要的构词形态是通过添加词缀的形式来表达,同时从词的结构和形态来看,高棉语的构词法可以分为三类,即单纯词构词法、合成词构词法和内部曲折法[1]。此外巴梵语的大量借用也对高棉语的构词形态产生了重要影响。

近年来，随着自然语言处理在各国的展开，高棉语的自然语言处理工作也受到越来越多研究者的重视。高棉语同许多其他亚洲语言一样是连续书写的，词与词之间没有明显的分隔符，因此对高棉语进行分词和词性标注研究很有必要。目前,这方面的研究已有少数机构开展了相关工作：由于采用最大匹配算法对高棉语进行分词准确率较低，且难以正确识别词库中没有的新词，因此蒋艳荣等人[2]提出采用改进的 Viterbi 算法，通过最优选择及剪枝操作来提高高棉语的分词效率，取得了一定的分词效率；Chea Sok Huor 等人[3]提出利用基于词语的二元文法模型和基于音节的二元文法模型对高棉语分词进行相关研究,实验结果表明基于词语的二元文法模型能取得相对较好的切分效果；Chenda NOU 等人[4]针对高棉语词性标注问题，提出结合基于规则和三元文法模型的混合方法对高棉语进行词性标注，在针对一个小型语料集的实验中取得了不错的效果。

高棉语在自产生以来，除经历了数次的文字改革外，还先后受到梵语、巴利语、法语、汉语、泰语、越南语等外来语言的影响，致使其具有词汇量丰富且构词复杂的特点。高棉语是由基本的辅音和元音等字符构成，基本字符首先根据特定的规则构成 KCC[5](Khmer Character Cluster, 字符簇)，然后由一个或多个 KCC 构成高棉语词素。就高棉语词语的意义和结构来看，高棉语语词语分为单纯词和合成词两类，单纯词指由一个独立词素构成的词，合成词指由两个或两个以上词素构成的词[6]。高棉语中合成词的比例明显高于单纯词，且其中有很多派生词，它们由独立词素加上一个词缀(辅助词素)构成。高棉语的词缀分为前缀和后缀，前缀可分为普通前缀、名词前缀和巴利语、梵语借词前缀。其中，名词前缀和巴利语、梵语借词前缀多是表示构词意义的辅助词素，而普通前缀则是表示语法意义的辅助词素。

针对高棉语复杂多变的构词特点,本文提出一种基于层叠条件随机场模型的高棉语自动分词及词性标注方法。该方法由三层条件随机场模型构成：第一层是分词模型，该模型以 KCC 为粒度，结合上下文信息与高棉语的构词特点，实现对高棉语文本的自动分词，然后将分词结果传递到第二层模型；第二层是分词结果修正模型，该模型以词语为粒度，结合上下文信息与高棉语命名实体的构成特点，实现对分词模型中切分错误的命名实体进行修正，然后将修正后的分词结果传递到第三层模型;第三层是词性标注模型,该模型以词语为粒度，结合上下文信息与高棉语丰富的词缀信息，实现对高棉语文本中的词语进行自动词性标注。经过这三层模型处理之后即得到最终的分词及词性标注结果。

2 基于层叠条件随机场的高棉语分词及词性标注方法

2.1 层叠条件随机场模型

高棉语的分词、词性标注等问题均可转化为序列标记问题，而条件随机场[7]是一种很好的进行序列标记的机器学习算法,它克服了 HMM 的独立性假设及最大熵模型的标记偏置等缺陷，为此本文采用条件随机场机器学习算法对高棉语的分词及词性标注问题进行建模。由于分词结果包含丰富的上下文信息以及高棉语词汇丰富的词缀信息都会对词性标注起到很好的指导作用，从而分词结果可以做为词性标注的输入，因此，需要引入多个层次的条件随机场模型用于高棉语的分词和词性标注。本文所提出的层叠条件随机场的模型共包含三层：第一层是分词模型，该模型以 KCC 为粒度，输入为经过分解 KCC 之后得到的 KCC 序列，结合上下文信息与高棉语的构词特点，实现对高棉语文本的自动分词，然后将分词结果传递到第二层模型；第二层是分词结果修正模型，之所以要加入第二层模型是由于在对第一层模

型的分词结果进行统计时发现，其中未登录词切分错误数约占整个分词错误数的 34%，经过分析发现未登录词切分错误主要由命名实体切分错误造成，而修正切分错误的命名实体不仅能修正上层模型中的分词结果，而且同时也能为下层词性标注模型提供更有效的支撑。该模型以词语(word)为粒度，结合上下文信息与高棉语命名实体的构成特点，实现对高棉语文本中的命名实体切分错误进行自动修正，然后将修正后的分词结果传递到第三层模型；第三层是词性标注模型，该模型同样以词语(word)为粒度，结合上下文信息与高棉语丰富的词缀信息，实现对柬文句子中的词语进行自动词性标注。整个三层模型的架构图如图 1 所示。

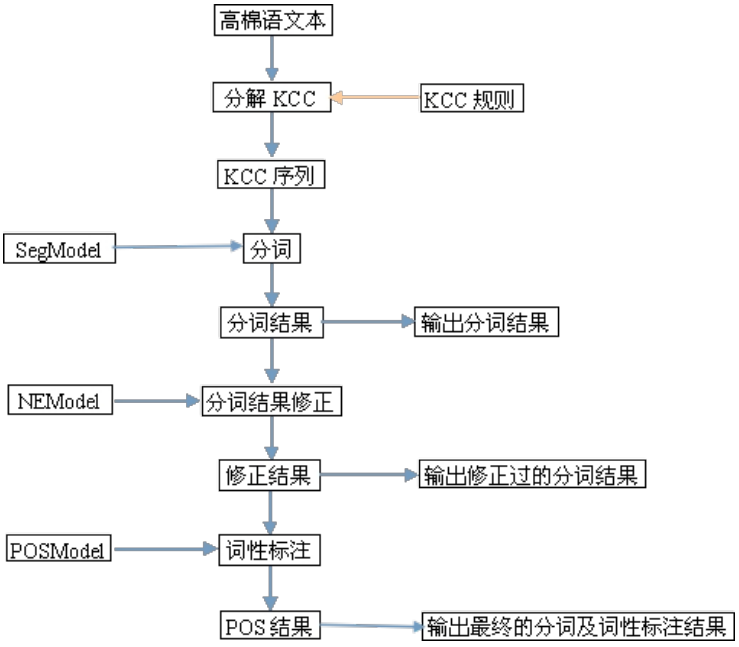

图 1 基于层叠条件随机场的高棉语分词及词性标注模型架构图

Figure 1 the architecture diagram of Khmer segmentation and POS tagging model based on CCFs

2.2 分词模型

通过对高棉语的语言学特征进行研究发现，高棉语是由基本的元音符号、辅音符号和一些其他符号构成，其中，这些符号首先按照特定规则构成具有固定读音的 KCC，该规则满足公式(1)[5]。

<C|I>+[<Robat|regshift>]+{COEUNG+<C+[regshift]|I+[regshift]>}+
[[<ZWJ|ZWNJ>]+V]+{S}+[ZWJ+COEUNG+<C|I>] (1)

其中，{}表示包含在其中的内容可出现 0-2 次，[]表示包含在其中的内容可出现 0-1 次，<x|y>表示选择 x 或选择 y，+表示把两边的部分连接起来。公式(1)中各表达式所表示字符的含义和 Unicode 码区间如表 1 所示，Unicode 码参照《Khmer Unicode Standard 6.2》[8]：

表 1 表达式-unicode 码对应表

Table 1 expression-unicode code correspondence table

符号	描述	Unicode 码区间	举例
C	33 个辅音符号	U+1780-U+17A2	ក、ខ、ឥ等
I	13 个独立元音符号	U+17A5-U+17B3	ឥ、ឦ、ឧ等
Robat	特殊符号，出现时表示 KCC 结束	U+17CC	◌៌

regshift	需要按 shift 按键才能输入字符	U+17C9、U+17CA	无
COEUNG	通过该符号可知道下一个字符为脚注	U+17D2	◌្
ZWJ	零宽的打印字符	U+200D	无
ZWNJ	零宽的非打印字符	U+200C	无
V	25 个元音字符	U+17B6-U+17C8	◌ា、◌ិ、◌ី等
S	12 个特殊符号	U+17C6-U+17D1	◌ះ、◌ុ 、◌ូ等

除公式(1)能匹配的 KCC 之外，对于高棉语文本中出现的国际标点符号、柬语标点符号、阿拉伯数字串、高棉语数字串、英文字符串等都同等当做 KCC 对待，将采用单独的正则表达式进行匹配识别。

通过上述规则可以将高棉语文本分解为包含多个 KCC 的序列。构建分词模型(SegModel)训练的特征模板定义如表 2 所示。

表 2 分词模型的特征模板

Table 2 feature template of SegModel

序号	特征类别	特征表示	特征描述
1	单 KCC	$K_i(-2\le i\le 2)$	K_i 表示单个的 KCC 作为特征，i 的取值表示相对当前 KCC 的偏移量，当 i=0 时即表示当前 KCC
2	两个 KCC	$K_iKj(-2\le i\le 1, j=i+1)$	K_iKj 表示相邻的两个 KCC 作为特征，i 和 j 的取值表示相对当前 KCC 的偏移量
3	三个 KCC	$K_iKjK_m(-2\le i\le 0, j=i+1,m=i+2)$	K_iKjK_m 表示相邻的三个 KCC 作为特征，i、j、m 的取值表示相对当前 KCC 的偏移量
4	标点	Boolean Mark(K_0)	当 K_0 是英文标点或柬语标点时特征值取 1，否则特征值取 0
5	数字	Boolean Number(K_0)	当 K_0 是阿拉伯数字串或柬语数字串时特征值取 1，否则特征值取 0
6	英文字符串	Boolean En_String(K_0)	当 K_0 是英文字符串时特征值取 1，否则特征值取 0

训练语料来源于 PLC(PAN Localization Cambodia)[1] 发布的公开语料集 Khmer Tagged Corpus[2](KCorpus)，KCorpus 是一个经过分词和词性标注处理的高棉语语料集，其中词性标签集采用 PLC 在 Khmer Part of Speech description[3] 中制定的词性标签集，通过高棉语语言学研究人员对 KCorpus 中出现的分词以及词性标注错误进行校正，得到一个包含 73127 词的高棉语语料集。首先需要将语料集中的每个词语进行 KCC 分解，并采用{B, M, E, S}四标记法对每个 KCC 进行标注；然后按照表 2 准备好分词特征模板；最后调用 CRF++工具包[9]，以分解过 KCC 的语料集和分词特征模板为输入，即可实现对 KCorpus 中的分词进行学习训练从而获得分词模型。构建分词模型的流程图如图 2 所示。

[1] http://www.panl10n.net/

[2] http:/www.panl10n.net/english/Outputs%20Phase%202/CCs/Cambodia/MoEYS/Software/2009/KhmerCorpus.zip

[3] http://www.PANL10n.net/wiki/PartOfSpeech

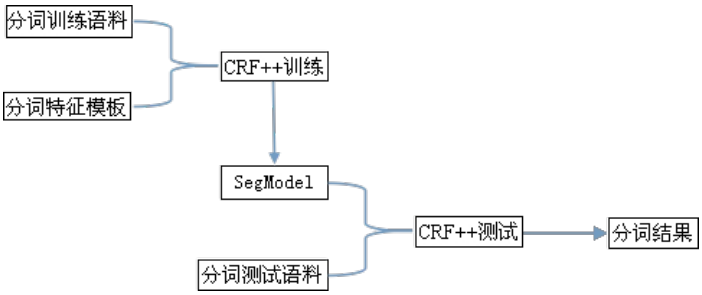

图 2 分词模型构建流程图

Figure 2 flow chart of SegModel building

2.3 分词结果修正模型

对采用 SegModel 进行开放分词测试的 40000 词中的错误进行统计，结果如表 3 所示，其中未登录词(Out of vocabulary，OOV)[10]被切分错误约占总错误数的 34%。若能对未登录词的切分错误进行修正，不仅可以进一步提高分词精度，同时也可以降低部分因错误传递而导致的词性标注的错误，进而可以从整体上提高分词以及词性标注的性能。基于上述考虑，故增加一层分词结果修正模型(NEModel)。

表 3 分词结果中的错误统计

Table 3 error statistics of segmentation results

类型		数量(个)	占总数比例(%)	错误率(%)
正确		37876	94.69	------
错误	OOV 切分错误	723	1.8075	34.0395
	交叉型歧义切分错误	221	0.5525	10.4049
	组合型歧义切分错误	1180	2.94	55.5556

通过对切分错误的未登录词进行分析发现，其中的绝大部分错误为人名、组织机构名、地名等命名实体的切分错误所造成，故此处的 NEModel 将针对高棉文的命名实体识别进行建模。由于对歧义切分错误的修正不能在层叠条件随机场的框架内进行，所以本文放弃对歧义切分的修正，仅对命名实体切分错误进行修正。经过对标注的高棉语命名实体的分析发现，当某些前缀和后缀出现时，连续出现的几个词成为命名实体的概率非常大，同时考虑上下文词汇特征，定义 NEModel 训练的特征模板如表 4 所示。

表 4 分词结果修正模型的特征模板

Table 4 feature template of NEModel

序号	特征类别	特征表示	特征描述
1	单字词	$W_i(-2\le i\le 2)$	W_i 表示单个的词语作为特征，i 的取值表示相对当前词的偏移量，当 i=0 时即表示当前词
2	两字词	$W_iW_j(-2\le i\le 1, j=i+1)$	W_iW_j 表示相邻的两个词语作为特征，i 和 j 的取值表示相对当前词的偏移量
3	三字词	$W_iW_jW_m(-2\le i\le 0, j=i+1, m=i+2)$	$W_iW_jW_m$ 表示相邻的三个词语作为特征，i、j、m 的取值表示相对当前词的偏移量
4	标点	Boolean Mark(W_0)	当 W_0 是英文标点或柬语标点时特征值取 1，否则特征值取 0
5	前缀	Begin_KCC(W_0)	取当前词 W_0 的第一个 KCC 为特征
6	后缀	End_KCC(W_0)	取当前词 W_0 的最后一个 KCC 为特征

训练语料来源于 PLC 发布的公开语料集 Khmer Tagged Corpus(KCorpus)。由于 KCorpus 中未对命名实体进行标注，所以本文首先利用人工方式将 KCorpus 中包含的命名实体进行

标注，标注时采用三名柬语语言学研究人员的投票方式以确定是否为实体，当投票方式仍不能确定是否为实体时将参考 English-Vietnamese Named Entity Guidelines[11]。经过对人工标注结果进行统计发现，KCorpus 语料集所包含的命名实体类别为人名、组织机构名、地名，故参照 English-Vietnamese Named Entity Guidelines 制定了如表 5 所示的高棉语命名实体规范以指导以后对高棉语词法分析的研究工作，随着研究升入此表有待进一步补充完善。对组成实体的各个词语进行标注时采用{B, M, E, S}四标记法，由于此处的命名实体识别是为了对上层的切分错误进行修正，故人标注时并不考虑实体的类别，仅标注是否为实体的组成部分。

表 5 高棉语命名实体规范

Table 5 Khmer named entity guidelines

实体类别	实体范畴
人名	名字、昵称、化名等
组织机构名	公司、组织、党派、运动队、团体、医院、酒店、学校、政府部门等
地名	海洋、湖、岛屿、河流、山、国家、省、市、区、街道、村庄、机场、公路、工厂等

在准备好 NEModel 训练语料和准备好特征模板之后，通过调用 CRF++工具包即可实现对 KCorpus 中的命名实体进行训练学习从而获取 NEModel。构建分词结果修正模型的流程图如图 3 所示。

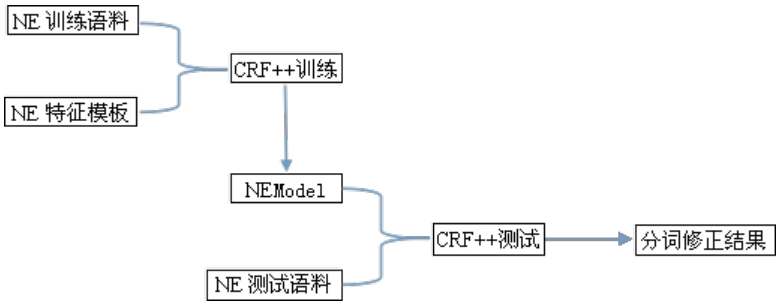

图 3 NEModel 构建流程图

Figure 3 flow chart of NEModel building

2.4 词性标注模型

高棉语文本经过分词和分词结果修正两个模型的处理后会得到更佳的切分结果，接下来需要对切分结果进行词性标注处理。据对大量的高棉语文本进行研究发现，高棉语的构词中包含丰富的词缀信息，高棉语构词附加成分承载着大量语法语义信息[6]。如果能对派生词缀的信息加以利用，设计适合柬文构词特性的特征模板，可有效提高高棉语词性标注的性能。由于高棉语词缀中既有表示语法意义(词性)的辅助词素，又有表示词义的辅助词素，且同一词缀可以表示多种词性。若采用规则的方法，不仅需要构造规则库，而且还需要考虑规则对语言现象的覆盖率和规则处理的正确率，通常这两种性能往往呈反比关系，且规则方法都有主观性，难以保证规则的一致性，适应性较差。故本文基于条件随机场模型在构建序列标记模型上的优势，采用基于条件随机场的高棉语词性标注方法，充分利用上下文信息和高棉语构词特点，在分析高棉语构词特点的基础上，融入高棉语词缀的信息特征，保留高棉语词缀中大量的语法、语义信息，其中词干的词性将按照上下文信息从比较权威的 Chuon Nath Khmer Dictionary[12]中提取，词性标记采用 PLC 在 Khmer Part-of-Speech Tagger[13]中发布的词性标记集。在训练语料上学习得到模型，不仅可以利用更多的上下文信息，而且还能够缓解

统计模型面临的数据稀疏问题。

获取词性标注模型(POSModel)训练用的特征模板定义如表 6 所示。

表 6 词性标注模型的特征模板

Table 6 feature template of POSModel

序号	特征类别	特征表示	特征描述
1	单字词	$W_i(-2\le i\le 2)$	W_i 表示单个的词语作为特征，i 的取值表示相对当前词的偏移量，当 i=0 时即表示当前词
2	两字词	$W_iW_j(-2\le i\le 1, j=i+1)$	W_iW_j 表示相邻的两个词语作为特征，i 和 j 的取值表示相对当前词的偏移量
3	三字词	$W_iW_jW_m(-2\le i\le 0, j=i+1, m=i+2)$	$W_iW_jW_m$ 表示相邻的三个词语作为特征，i、j、m 的取值表示相对当前词的偏移量
4	词+词性	$W_iP_i(-2\le i\le 2)$	W_iP_i 表示取词及其词性作为特征，i 的取值表示相对当前词的偏移量，当 i=0 时即表示当前词
5	标点	Boolean Mark(W_0)	当 W_0 是英文标点或柬语标点时特征值取 1，否则特征值取 0
6	第一个 KCC	First_KCC(W_0)	取当前词 W_0 的第一个 KCC 为特征
7	第二个 KCC	Second_KCC(W_0)	取当前词 W_0 的第二个 KCC 为特征，无时特征值取 0
8	第三个 KCC	Third_KCC(W_0)	取当前词 W_0 的第三个 KCC 为特征，无时特征值取 0
9	第一个 KCC	End_KCC(W_0)	取当前词 W_0 的最后一个 KCC 为特征
10	第二个 KCC	Pre_End_KCC(W_0)	取当前词 W_0 的倒数第二个 KCC 为特征，无时特征值取 0
11	词干	Real_Word(W_0)	提取 W_0(当前词)的词干
12	词干的词性	P(Real_Word(W_0))	取 W_0(当前词)的词干的词性

训练语料来源于 PLC 发布的公开语料集 Khmer Tagged Corpus(KCorpus)。首先采用人工的方式对 KCorpus 语料集进行校对，对其中出现的词性标注错误进行修正，词性标签集参照 Khmer Part-of-Speech Tagger。在准备好训练语料和特征模板后，通过调用 CRF++工具包即可实现对 KCorpus 中的词性标注进行学习训练从而获取 POSModel。构建词性标注模型的流程图如图 4 所示。

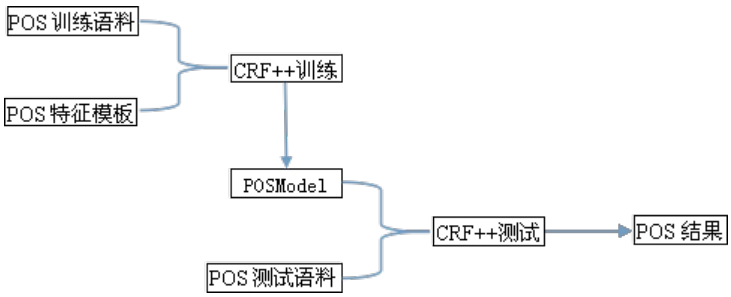

图 4 词性标注模型的构建流程图

Figure 4 flow chart of POSModel building

3 实验结果与分析

为了能客观评价基于层叠条件随机场模型(CCRFs)的高棉语分词及词性标注模型的效

果，我们做了两组对比实验，第一组实验是单层模型(CRFs)、两层模型(CCRFs，分词模型+分词结果修正模型)与最大匹配算法[2]、MViterbi 算法[2]以及基于词语的二元文法模型[3]的高棉语分词开放测试实验，第二组实验是三层模型(CCRFs，分词模型+分词结果修正模型+词性标注模型)与基于变换的方法[4]的高棉语词性标注开放测试实验。本文做开放测试实验的语料来源于从柬埔寨新闻网[1]上爬取的自 2012 年 10 月至 2013 年 7 月的新闻文本，涵盖艺术、娱乐、体育、健康、经济、国内以及国际等七大新闻板块，约 10M(50 万词)。

为了更具体地说明采用三层模型(CCRFs)对高棉语句子进行 KCC 分解、分词、分词结果修正以及词性标注的过程，下面我们以柬语句子ខ្ញុំស្រឡាញ់ប្រទេសចិន(译：我爱中国)为例进行说明。原句子经过 KCC 分解后变成ខ្ញុំ ស្រ ឡា ញ់ ប្រ ទេ ស ចិ ន(各 KCC 之间以空格分隔)；经过分词模型分词后变为ខ្ញុំ / ស្រឡាញ់ / ប្រទេស / ចិន(译：我/爱/国家/中国，各词语之间以“/”分隔)；分词结果中ប្រទេស(国家)和ចិន(中国)能组合为命名实体ប្រទេសចិន(中国)，因此需要修正分词结果，经过分词结果修正模型修正后变为ខ្ញុំ / ស្រឡាញ់ / ប្រទេសចិន(译：我/爱/中国)；经过词性标注模型标注后变为ខ្ញុំ/PRP ស្រឡាញ់/V ប្រទេសចិន/NNP(各词语之间以空格分隔，“/”的前后分别是词语以及该词对应的词性)。

测试结果采用常用的三个评测指标：准确率(P)、召回率(R)和 F 值(F)[14]。实验结果如表 7，表 8 所示。

表 7 最大匹配算法、MViterbi 算法、基于词语的二元文法模型、单层模型(CRFs)、两层模型(CCRFs)的高棉语分词实验结果比较

Table 7 the results contrast of MM、MViterbi、2-gram、CRFs、CCRFs

测试类型	准确率(%)	召回率(%)	F1 值(%)
最大匹配算法	72.26	------	------
MViterbi 算法	88.17	------	------
基于词语的二元文法模型	91.562	92.138	91.849
单层模型(CRFs)	94.69	95.06	93.82
两层模型(CCRFs)	96.02	96.49	96.25

从表 7 的分词对比实验结果可以看出，两层模型(CCRFs，分词模型+分词结果修正模型)的分词效果相对于单层模型(CRFs)以及其他方法，其正确率、召回率、F1 值都有所提高。尤其是相对于前三种方法中效果最好的基于词语的二元文法模型来说，两层模型的准确率提高了近 4.5 个百分点，从而说明本文提出的采用层叠条件随机场(CCRFs)来解决高棉语的分词问题是非常有效的。另外，相对于单层模型(CRFs)而言，两层模型的准确率在其基础之上也提高了近 1.5 个百分点，这也说明加入分词结果修正模型是有必要的，能将分词模型中切分错分的命名实体进行修正，从而提高分词结果的准确率，避免将部分错误传递给下层模型从而影响最终的词性标注效果。

表 8 基于变换的方法与三层模型(CCRFs)的词性标注实验结果比较

Table 8 the results contrast of transformation-based approach and CCRFs

[1] http://www.dap-news.com

测试类型	准确率(%)	召回率(%)	F1 值(%)
基于变换的方法	91.96	------	------
三层模型(CCRFs)	95.44	94.57	90.28

从表 8 的词性标注对比实验结果可以看出，三层模型(CCRFs，分词模型+分词结果修正模型+词性标注模型)的词性标注效果相对于基于变换的方法，其正确率提高了近 3.5 个百分点，说明本文提出的基于三层的层叠条件随机场模型的高棉语词性标注方法是行之有效的。

4 结论

本文针对高棉语分词和词性标注问题，提出了一种基于三层的层叠条件随机场模型高棉语分词和词性标注方法，在充分考虑上下文信息对高棉语分词和词性的影响外，通过对大量高棉语文本进行总结归纳，得到一些可以利用的高棉语的自身特点，并设计有效的特征模板。通过跟单层模型(CRFs)、两层模型(CCRFs，分词模型+分词结果修正模型)、最大匹配算法、MViterbi 算法、基于词语的二元文法模型以及基于变换的方法等的开放测试结果对比，结果表明本文提出的方法取得了不错的效果。但在实验中发现两个问题：一是在对第一层分词模型的开放测试进行统计后发现，组合型歧义切分错误占到错误总数的 54.3%，约为排在第二的命名实体切分错误的两倍，虽然对命名实体的切分错误进行了修正，但通过实验可以看出对最终的准确率提高并不是太明显，若能解决组合型歧义切分错误，将会更大的提高最终的分词准确率；二是由于高棉语构词的特殊性，分词之前需要先对高棉语文本进行分解 KCC 处理，这个处理过程增加了系统的资源开销，也增加了系统的时间开销，同时三层模型虽然提高了分词和词性标注的精度，但这在一定程度上是以损失系统的时间效率为代价的，另外经统计中得出分词速度约为 32.65kb/s，仍有很大的提升空间。下一步的工作将在修正组合型歧义切分的问题上展开，以期能进一步提高高棉语分词和词性标注的效果。

参考文献

[1] 莫源源. 高棉语的构词方式及其语法功能[1][J]. 无线音乐·教育前沿, 2012 (10).

[2] 蒋艳荣, 刘习文, 陈耿涛. 基于 Viterbi 改进算法的高棉语分词研究[J]. Computer Engineering, 2011, 37(15).

[3] Huor C S, Rithy T, Hemy R P, et al. Word Bigram Vs Orthographic Syllable Bigram in Khmer Word Segmentation[J]. PAN Localization Working Papers, 2004, 2007.

[4] Nou C, Kameyama W. Khmer POS Tagger: A Transformation-based Approach with Hybrid Unknown Word Handling[C]//Semantic Computing, 2007. ICSC 2007. International Conference on. IEEE, 2007: 482-492.

[5] Huor C S, Hemy R P, Navy V. Detection and Correction of Homophonous Error Word for Khmer Language[J]. Ref. No. PANL10n/Admn/RR.

[6]肯素(柬埔寨).高棉语法[M].柬埔寨皇家科学院出版社,2007.5.

[7]Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[8]Unicode Consortium. The Unicode Standard, Version 6.2. 0 (Mountain View, CA, The Unicode Consortium, 2012, ISBN 978-1-936213-07-8)[J]. 2012.

[9]Kudo T. CRF++: Yet another CRF toolkit[J]. Software available at http://crfpp. sourceforge. net, 2005.

[10]Bazzi I. Modelling out-of-vocabulary words for robust speech recognition[D]. Massachusetts Institute of Technology, 2002.

[11]Quoc Hung Ngo.English-Vietnamese Named Entity Guidelines,Version 1.0,English January 24,2012.

[12]CHUON NATH, Dictionnaire Cambodgien, Edition de L'Institut Bouddhique, Phnom Penh, 1967.
[13]Chena N O U. Khmer Part-of-Speech Tagging[J]. Global Information and Telecommunication Studies, WASEDA UNIVERSITY.
[14]郭剑毅, 薛征山, 余正涛, 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.

作者联系方式：

姓名	地址	邮编	电话	电子邮箱
潘华山	云南省昆明市昆明理工大学呈贡校区信自楼 540 室	650500	15887811064	793822968@qq.com
严馨	云南省昆明市昆明理工大学呈贡校区信自楼 520 室	650500	18206713208	kg_yanxin@sina.com