Computer Science and Application 计算机科学与应用, 2018, 8(11), 1682-1688
Published Online November 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.811185

Hans汉斯

Burmese Segmentation Methods and Its Implementation

Chang'e Ma, Jian Yang*

School of Information Science and Engineering, Yunnan University, Kunming Yunnan
Email: *jianyang@ynu.edu.cn

Received: Oct. 21st, 2018; accepted: Nov. 6th, 2018; published: Nov. 13th, 2018

Abstract

Unlike English and other western languages, there are no delimiters to mark word boundaries in Burmese. Therefore, word segmentation is an important part in the realization of Burmese speech synthesis. Through manually word segmentation by Burmese experts, we have constructed a Burmese text database containing 5000 sentences as experimental data of this paper. The CRF-based word segmentation method is compared with the FMM-based word segmentation method. The performance of word segmentation method was evaluated with confidence, precision and speed of segmentation. In this experiment, the confidence of the Burmese word segmentation the CRF-based and FMM-based was 94.1% and 84.3%, respectively, and the F values were 93.8% and 82.9%, respectively. It shows that the CRF method can be applied to Burmese word segmentation with better effect. We believe that this method meets the requirements for the development of the Burmese speech synthesis system.

Keywords

Burmese, Segmentation, CRF, FMM

缅甸语分词方法及其实现

马昌娥，杨　鉴*

云南大学信息学院，云南　昆明
Email: *jianyang@ynu.edu.cn

收稿日期：2018年10月21日；录用日期：2018年11月6日；发布日期：2018年11月13日

摘　要

缅甸语与英语以及其它西方语言不同，它的词之间没有明显的边界，开发缅甸语的语音合成系统时，分

*通讯作者。

文章引用: 马昌娥, 杨鉴. 缅甸语分词方法及其实现[J]. 计算机科学与应用, 2018, 8(11): 1682-1688.
DOI: 10.12677/csa.2018.811185

词是其中的一个重要环节。我们从大约600 M的原始语料库中选取5000个完整句子，由缅语专家人工分词以后作为该文的实验数据集。本文对比了基于条件随机场(CRF)的缅语分词方法与基于正向最大匹配算法(FMM)的缅语分词方法，并用置信度、分词精度和分词速度评估分词性能。在本次实验中，基于CRF与FMM的缅语分词结果中置信度分别可达94.1%和84.3%，F-值分别可达93.8%和82.9%。表明，应用CRF方法实现缅语分词的效果更好，且该方法可满足开发缅语语音合成系统的要求。

关键词

缅甸语，分词，条件随机场，正向最大匹配算法

1. 引言

近些年，缅语分词逐渐受到研究人员的关注，但是相比中文分词的成果，缅语分词相关的研究还需要继续努力。目前，主要的分词方法有三类：1) 基于词典匹配的方法，如果发现文本中出现字符串的某一子串与词典中的词条相同，就算是匹配成功。它的优点是速度快，实现简单，缺点是无法正确识别未登录词。2) 基于统计的分词方法，上下文中，相邻的字出现的次数越多，就越可能构成一个词。因此字与字相邻出现的概率或频率能很好的反映词的可信度。它的优点是分词准确性较高，但分词速度较慢。3) 基于规则的分词，通过模拟人对句子的理解，达到识别词的效果。基本思想是语义分析，句法分析，利用句法信息和语义信息文本进行分析，它的优点是分词准确性高，但算法实现复杂，实现难度较大。对比以上三种分词方法，由于缅语的复杂性和相对匮乏性，暂时不考虑基于规则的缅语分词方法。

关于缅语分词方面的研究，Thet T T 和 Na J C 等人曾提出了一种基于规则的音节划分与基于字典的音节合并的缅语分词方式[1]，Mo H M 和 Nwet K T 等人提出了一种基于 CRF 的缅语人名识别方法[2]，Phyu M L 和 Hashimoto K.等人提出了一种基于 CRF 和特征聚类的缅语分词方式[3]。本文以开发缅语语音合成系统为目的，分析和比较了基于 CRF 和 FMM 的分词效果。为了能在缅甸语分词过程中使用 CRF 模型，我们构建了一个人工分词语料库作为训练集，通过 0.58 版本的 CRF++工具包[4]在 Linux 平台上实现缅语了分词系统。

本文的其余部分安排如下：第一部分，介绍了缅甸语的特征。第二部分，缅语分词方法的介绍。第三部分，实验结果以及分析。第四部分，总结与展望。

2. 缅甸语简介

缅甸语属于汉藏语系藏缅语族，它是缅甸社会主义联邦共和国的通用语，大约有 5400 万人使用这一种重要的语言。缅语不同于英文以及欧洲其他的一些语言，它与中文相似，都是大字符集上的连续子串，且没有明显的词边界。现代缅甸文字是一种拼音文字，书写时从左到右。从字型来看，好像是用大小不同的圆圈拼接、套叠而成。它在东方和世界文字之林中，可以说是别具一格，不同于方块形的汉子表意，缅文表音。缅甸语有 33 个基本辅音字母，7 个独立的元音字符，7 个非独立元音字符，4 个基本介音，10 个数字符号，2 个标点符号，详见文献[5]。下面是一个缅语句子划分为音节和音节拼接为词的实例，音节之间和词之间均用“+”连接。

句子：ကျွန်တော်သည် ကျောင်းသားတစ်ယောက် ဖြစ်သည်။
音节：ကျွန်+တော်+သည်+ကျောင်း+သား+တစ်+ယောက်+ဖြစ်+သည်+။
词：ကျွန်တော်+သည်+ကျောင်းသား+တစ်ယောက်+ဖြစ်သည်+။
中译：我是一名学生。

3. 缅语分词方式

3.1. 基于 CRF 的分词

条件随机场模型(Conditional Random Fields，简称 CRF，或 CRFs)是基于隐马尔科夫模型和最大熵模型提出的一种判别式概率无向图学习模型[6]，也是一种基于统计序列标注和分割的分词方法[7]，它将缅语分词问题转化为一个序列标注问题，即将词语开始至结束字符分别标记，之后根据标记完成对句子分词。相比隐马尔可夫模型和最大熵模型，CRF 没有隐马尔可夫模型那样严格的独立性假设(观测值之间相互独立)，因而可以容纳上下文信息，同时 CRF 还克服了最大熵模型标记偏置的缺点。

输入数据的序列标注即为每一个输入序列分配一个标注符号，对于分词问题，$X=(x_1,x_2,\cdots,x_T)$ 是一个长度为 T 的输入数据序列，$Y=(y_1,y_2,\cdots,y_T)$ 作为对应的标注序列。在给定一个输入序列的情况下，设 $p(y|x)$ 为线性链条件随机场，则在随机变量 X 取值为 x 的条件下，随机变量 Y 的取值为 y 的条件概率具有如下形式，如公式(1)：

$$p(y|x)=\frac{1}{Z_X}\exp\left(\sum_{1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,x,t)\right) \tag{1}$$

其中，Z_x 是规范化因子，$f_k(y_{t-1},y_t,x,t)$ 是一个任意的特征函数，λ_k 是与特征函数相关的权重。条件随机场的预测问题是给定条件概率 $p(y|x)$ 和输入序列 x，求条件概率最大的标注序列 y^*，即对输入序列进行标注。条件随机场的预测算法是著名的维特比算法，如公式(2)：

$$y^*=\arg\max_{y} P(y|x) \tag{2}$$

3.1.1. 标注方式

条件随机场的缅文分词思想类似于中文分词思想，就是通过分析大量的缅语文本语料库去计算一个音节在字符串中的位置的可能性。利用条件随机场来进行缅文分词就是一个字符标注的过程，缅文文本的标注以一个音节为基本单位，缅甸语的词是由单个音节或多个音节构成，统计 5000 个缅语句子，共计 124,809 个词，去除重复的词语后得到 23,098 个词，分别统计去重复前和去重复后缅语的词长分布，如表 1 所示。

Table 1. Situation of Burmese syllable in words
表 1. 缅语词长统计表

音节类型	所占百分比(去重前)	所占百分比(去重后)	例子
单音节	49.0%	17.0%	ပြေး (跑)
双音节	26.8%	30.5%	တိုက်ခိုက် (攻击)
三音节	16.9%	29.8%	လုပ်ငန်းရှင် (企业家)
四音节	5.3%	14.7%	အမျိုးအစား (种类)
四音节以上	2.0%	8.0%	ကြောက်စရာကောင်းသော (可怕的)

由表 1 可知，去重前缅甸语中的词语的大多数为单音节词、双音节词和三音节词，四音节词及以上相对较少，统计得到缅语的词长均值为 1.86 个音节，去重后双音节词和三音节词所占比例较大。本文使用“S”“B”“M”“E”来分别表示单音节词、该音节出现在词的头部、该音节出现在在一个词的中间位置和该音节出现在一个词的词尾位置。缅甸语文本“ဒီနေ့ရာသီဥတုအရမ်းသာယာတယ်။”标注过程如图 1 所示，图 1 中过程①将句子分为五个词，过程②表示将每一个词划分为对应的音节并进行标注。

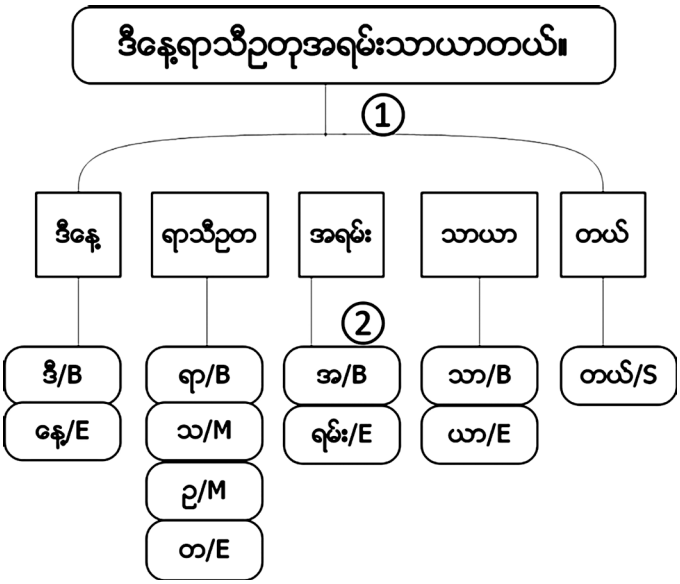

Figure 1. Burmese text tagging
图 1. 缅文文本标注

3.1.2. 特征模板

CRF 中一共存在两种模板：U-gram 和 B-gram，U-gram 为一元模板，表示只与当前位置对应的标签相关的特征，B-gram 二元模板，表示前一个位置和当前位置对应的标签相关的特征，当类别数很大的时候，这种类型会产生许多可区分的特征，这将会导致训练和测试的效率都很低下，本文选择的是 U-gram 模板。对字符串标注前还要进行特征的选择，虽然条件随机场模型中在理论上可以允许各种长度各种样式的特征，但是如果我们选择的特征过多过长会使得整个系统在运行算法时变得速度很慢，所以通常我们选取特征只在有限的上下文本范围内进行选择。特征模板格式：%x [row, col]。x 可取 U 或 B，对应两种类型。方括号里的编号用于标定特征来源，row 表示相对当前位置的行，0 即是当前行；col 对应训练文件中的列，这里只使用第 1 列(编号 0)。本文选择距离为 2 的特征模板，其具体含义如表 2 所示。

3.2. 基于 FMM 的分词

本文对基于词典的分词算法采用正向最大匹配(Forward Maximum Matching，简称 FMM)，其主要原理都是切分出单音节串，然后和词库进行比对，如果是一个词就记录下来，否则通过增加或者减少一个音节，继续比较，一直还剩下一个音节则终止，如果该单音节串无法切分，则作为未登录处理。同时，缅语文本中存在四音格结构[8] [9]，缅甸语四音格词的类型为 ABCD、ABAC、ABCB、AABB、ABCC、ABBC、ABCA、ABAA、ABAB、AAAA、AABC，共 11 种。其中，数量最多的是 ABCD 类型，最少的是 ABAA 类型。本文用到 AAAA、AABB、ABAA、ABAB 四种四音格词，即若音节列表中连续出现四

个音节满足以上四种四音格词结构，则将这四个音节重划为一个词。

词典来自 Technomaion Studios 公司编著的缅甸语-英语词典，共有词条 26,601 个，为了充分和合理的使用词典，我们提出了基于 Hash 的词典设计方式从而能够更快更有效率的分词。

Table 2. Feature Template
表 2. 特征模板说明

特征	含义
%U [−2, 0]	当前字与当前字向前第二个字的联系
%U [−1, 0]	当前字与当前字向前第一个字的联系
%U [0, 0]	当前字与当前字的联系
%U [1, 0]	当前字与当前字向后第一个字的联系
%U [2, 0]	当前字与当前字向后第二个字的联系
%U [−2, 0]/%U [−1, 0]/%U [0, 0]	当前字与当前字及当前字向前第一、二个字的联系
%U [−1, 0]/%U [0, 0]/%U [1, 0]	当前字与当前字、当前字向前第一个字、当前字向后第二个字的联系
%U [0, 0]/%U [1, 0]/%U [2, 0]	当前字与当前字及当前字向后第一、二个字的联系
%U [−1, 0]/%U [0, 0]	当前字与当前字及当前字向前第一个字的联系
%U [0, 0]/%U [1, 0]	当前字与当前字及当前字向后第一个字的联系

4. 实验方案与结果分析

4.1. 实验数据库的构建以及分析

本文实验数据库的构建过程为：

1) 利用爬虫软件火车采集器在缅甸语网站上抓取大约 600 MB 的原始语料库，内容涵盖了娱乐、体育、生活、商业、宗教、新闻等各方面。

2) 原始语料经过去重复和去非法字符后得到文本语料库，从文本语料库中随机挑选出 5000 句作为缅语分词实验文本集，经过音节划分之后，采用基于词典的分词方法对数据进行初步划分；

3) 将基于词典的分词结果交给两位缅甸语专家进行校对以及改正，缅语专家按语义对语料分词，将 5000 个缅语句子平均地分为 25 个文档，编号为#1，#2，#3，……，#25。A 专家和 B 专家各自需要完成的文档如表 3 所示，#1 至#5 号文档由专家 A 完成后得到 TextSet1 文本集，包含 5 个文档，编号#A1~#A5，由专家 B 完成后得到 TextSet2 文本集，包含 5 个文档，编号#B1~#B5，#6 至#15 号文档由专家 A 完成后得到 TextSet3 文本集，#16~#25 号文档由专家 B 完成后得到 TextSet4 文本集。

本文根据置信度来衡量实验结果的真实性，假设 A 专家、B 专家对同一个文档完成分词以后得到的两个文档#A_i和#B_i (I = 1, 2, 3, 4, 5)中词的数量分别表示为$N(A_i)$、$N(B_i)$，其中相同的词的数量表示为$M(A_iB_i)$。那么 B_i 相对于 A_i 的置信度表示为公式(3)：

$$C\left(B_i, A_i\right) = P\left(B_i \mid A_i\right) = M\left(A_iB_i\right) \big/ N\left(A_i\right) \tag{3}$$

若将 TextSet1 文本集作为标准分词数据，那么 TextSet2 文本集相对于 TextSet1 文本集的置信度为 $C(B_i, A_i)$，阈值 $C1 = \min\{C(B_i, A_i)\}$。若将 TextSet2 文本集作为标准分词数据，那么 TextSet1 文本集相对于 TextSet2 文本集的置信度为 $C(A_i, B_i)$，阈值 $C2 = \min\{C(A_i, B_i)\}$，其中 i = 1, 2, 3, 4, 5，置信度的计算结

果如表 4 所示。

根据表 5，$C(B_i, A_i)$的最小值为 0.875，则基于文本集 TextSet1 时，选定阈值为 0.875。同理，基于文本集 TextSet2 时，选定阈值为 0.906。

另外，分词精度和分词速度是评判一个分词系统性能好坏的两个重要方面，分词速度主要就是对待缅语分词，机器运算的处理速度，在机器硬件条件一样的前提下，分词速度越快越好。分词精度则主要是指的分词能够达到的准确程度，这是评价一个分词系统的最核心的标准。本文对于分词精度将选用召回率 R、准确率 P 以及 F-值作为评判标准[10]，其定义如公式(4)，(5)，(6)。其中，F-值能够综合衡量准确率和召回率，是 P 和 R 之间的权重比，令 $\beta=1$，则对待 P 和 R 的权重相同。A 表示测试文本中切分正确的词总数，B 表示测试文本切分错误的词总数，C 表示在标准文本中属于正确答案但是在测试文本中没有被识别出来的词总数。

$$召回率(R)=\frac{测试文本中切分正确的词总数}{标准文本中的词总数}\times 100\%=\frac{A}{A+C}\times 100\% \tag{4}$$

$$准确率(P)=\frac{测试文本中切分正确的词总数}{测试文本中的词总数}\times 100\%=\frac{A}{A+B}\times 100\% \tag{5}$$

$$F=\frac{P\times R\times\left(\beta^2+1\right)}{R+\beta^2\times P}\times 100\% \tag{6}$$

Table 3. Segmentation distribution by expert A and B
表 3. 专家 A 和专家 B 的分词文档

文档号 \ 专家号	专家 A	专家 B
#1~#5	√	√
#6~#15	√	
#16~#25		√

Table 4. Results of the Confidence calculation
表 4. 置信度计算结果

文档	$N(A_i)$	$N(B_i)$	$M(A_iB_i)$	$C(B_i, A_i)$	$C(A_i, B_i)$
#1	6217	6105	5537	89.1%	90.7%
#2	6452	6259	5684	88.1%	90.8%
#3	6567	6402	5912	90.0%	92.3%
#4	6098	5968	5437	89.2%	91.1%
#5	6758	6523	5910	87.5%	90.6%

4.2. 分词实验结果对比及分析

将 TextSet3 文本集和 TextSet4 文本集作为 CRF 分词系统中的训练数据，#1~#5 号文档作为测试数据进行 CRF 分词，同时也采用基于词典的分词方式，为了充分和合理的使用词典，我们提出了基于 Hash 的词典设计方式从而能够更快更有效率的分词，基于词典的算法采用正向最大匹配算法(FMM)，分别将两种分词结果与 TextSet1 文本集和 TextSet2 文本集比较后，得到表 5 所示结果。

Table 5. Segmentation results of CRF and FMM methods
表 5. CRF 和 FMM 分词结果

系统	标准文本集	置信度	召回率	准确率	F-值	分词速度
CRF	TextSet1	92.3%	92.3%	92.7%	92.5%	138KB/s
	TextSet2	94.1%	94.1%	93.6%	93.8%	
FMM	TextSet1	84.3%	84.3%	80.9%	82.2%	186KB/s
	TextSet2	82.0%	82.0%	83.9%	82.9%	

根据表 5 可知，不论是基于文本集 TextSet1 还是基于文本集 TextSet2，采用 CRF 分词得到的置信度大于阈值，采用 FMM 分词得到的置信度小于阈值，并且 CRF 分词的 F-值相比 FMM 分词提高了 10 个百分点以上。而在机器的硬件条件相同的情况下，对同样大小的文档分词，CRF 的分词速度相比 FMM 分词的速度慢，未来将考虑统计和词典结合的分词方法以提升速度和精度。

5. 结语

通过置信度、分词精度和分词速度比较了基于 CRF 和基于 FMM 的缅语分词方法，实验结果表明，前者对缅语分词具有较好的效果，可以将基于 CRF 的分词方式初步应用于缅语的分词。但是在分词速度上后者更占优势，因此，需要进一步考虑更完善的系统同时提升速度和精度。

基金项目

本文获得国家自然科学基金项目(61262068)资助。

参考文献

[1] Thet, T.T., Na, J.C. and Ko, W.K. (2008) Word Segmentation for the Myanmar Language. *Journal of Information Science*, **34**, 688-704. https://doi.org/10.1177/0165551507086258

[2] Mo, H.M., Nwet, K.T. and Soe, K.M. (2016) CRF-Based Named Entity Recognition for Myanmar Language. Genetic and Evolutionary Computing. Springer International Publishing, 204-211.

[3] Phyu, M.L. and Hashimoto, K. (2017) Burmese Word Segmentation with Character Clustering and CRFs. *International Joint Conference on Computer Science and Software Engineering*, 1-6.

[4] CRF toolkit. https://github.com/phychaos/pycrfpp

[5] Chang'e, M. and Jian, Y. (2018) Burmese Word Segmentation Method and Implementation Based on CRF. *International Conference on Asian Language Processing*, 11.

[6] Sutton, C. and McCallum, A. (2011) An Introduction to Conditional Random Fields. Now Publishers Inc, Hanover, MA, 40: 267-373.

[7] Lafferty, J.D., Mccallum, A. and Pereira, F.C.N. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence TextSet. *Proceedings of ICML*.

[8] 高凯敏. 汉缅语四音格词比较研究[D]: [硕士学位论文]. 昆明: 云南师范大学, 2007.

[9] 荣晶. 藏缅语族的四音格形式[J]. 云南民族大学学报: 哲学社会科学版, 2003, 20(4): 215-219.

[10] Sproat, R. and Emerson, T. (2003) The First International Chinese Word Segmentation Bakeoff. *Proceedings of the Second Sighan Workshop on Chinese Language Processing*, **17**, 133-143. https://doi.org/10.3115/1119250.1119269

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org