傣文自动分词系统的设计与实现[i]

高廷丽[1]，戴红亮[2]
[1]中国科学院自动化研究所模式识别国家重点实验室，北京
tingli.gao@nlpr.ia.ac.cn
[2]教育部语言文字应用研究所，北京
daihongliang9119@126.com

摘 要：傣文自动分词是傣文信息处理中的基础工作，是后续进行傣文输入法开发、傣文自动机器翻译系统开发、傣文文本信息抽取等傣文信息处理的基础，受限于傣语语料库技术，傣文自然语言处理技术较为薄弱。本文首先对傣文特点进行了分析，并在此基础上构建了傣文语料库,同时将中文分词方法应用到傣文中，结合傣文自身的特点，设计了一个基于音节序列标注的傣文分词系统，经过实验，该分词系统达到了 95.58%的综合评价值。

关键词:傣文；分词；CRF；绝对切分词

中图分类号：TP391　　文献标识码：A

DAIWEN WORD SEGMENTATION SYSTEM DESIGN AND IMPLEMENTATION

Abstract: Daiwen word segmentation is the basis for Daiwen information processing work. It's the basic work for Daiwen input method, Daiwen machine translation system development, daiwen text information extraction and other information processing words. Limited by Daiwen corpus technology, Daiwen natural language processing technology is relatively weak. This paper first analyzes the characteristics of Daiwen, and on this basis, build a Daiwen corpus, then, applied Chinese word segmentation method to Daiwen segmentation, combined with its own characteristics, Designed an Daiwen word segmentation system based on the sequence annotation. Through experiments, the segmentation system has reached a comprehensive appraisal 95.58%.

Key words: Daiwen; Segmentation; CRF; Absolute segmentation word

1　引言

随着计算机技术和互联网技术的发展，积累了大规模的中文语言资源，为中文分词、文本分类、输入法等中文信息处理技术提供了良好的研究基础。目前，中文分词技术已经非常成熟，准确率和召回率达到了较高水平。西双版纳傣文做为一种少数民族语言文字，与中文有较大差别,目前虽然有部分傣文语料，但都没有经过加工处理，无法用于后续研究开发。受限于语料库技术，傣文的自然语言处理技术也较为薄弱，目前还没有较好的傣文分词系统。本文的工作主要有两部分：一是建立一个能用于后续傣文信息处理的傣文语料库；二是充分结合傣文自身特点，将中文分词方法应用到傣文中，研制一个高性能的傣文分词系统，为傣文信息处理提供基础。

2 西双版纳傣文及特点

傣族有五种不同形体的文字，即西双版纳傣文（也称傣泐文或经典傣文）、德宏傣文（也称傣那文）、傣绷文、金平傣文、新平傣文。其中西双版纳傣文和德宏傣文使用广泛，留下了大量文献，傣绷文也留有少量文献，金平傣文和新平傣文文献较罕见。

西双版纳傣文当地傣语称为“多泐 ki VAh”，ki 为“文字”的意思，VAh 为西双版纳傣族自称，意为傣泐人使用的文字。但在泰国、缅甸北部和老挝以及我国其他傣族地区则多称西双版纳傣文为“多塔姆 ki OF”，OF 意为佛经，连起来就是佛经文字。傣文源于印度的婆罗米字母，是随着小乘佛教的传播进入傣族地区的。婆罗米字母在演变过程中分成三个大的子系统，其中阿萨姆文、他加禄文、那加利文、天成体梵文字母和藏文与梵文较为接近，其他文字分成南北两系，南系文字主要有泰米尔文，爪哇文等，北系文字是巴利文文字及其衍生文字（主要包括缅甸文、僧伽罗文、波罗马特文、高棉文、泰国文、老挝文以及我国的傣文）。西双版纳傣文有新老傣文之别。老傣泐文的组合结构比较复杂，一般是辅音字母居中，顶格书写，元音字母或韵母符号以及声调符号分布于辅音字母的上下左右，多数字词以辅音为中心单向或双向向外扩散，有的字词以辅音为中心向四周扩散。有时前一字母和后一字母又交叉组合，环环相扣。1954 年，西双版纳通过了“西双版纳傣文改进方案”，在继承了老傣文字形和声母分高低音组的基础上，对老傣文做了一下改进：

1.以西双版纳傣族自治州政府所在地景洪语音为标准音，依据语音实际情况，增添了老傣泐文没有的辅音字母；

2.依据一声母、一韵母、一声调一种写法的原则，选择老傣文中使用频率高的字母作为声母书写形式，改变了大部分元音和韵尾的写法，重新规定了声调的写法；

3.改变了老傣泐文以声母为核心的书写形态，将傣文声韵母线性化。

傣语属于汉藏语系壮侗语组壮傣语支，是一种有声调的单音节文字，音节类型与汉语相近，傣语一个词通常由一个或多个音节组成，因此傣文在信息处理时，也需要像汉语一样进行分词。

3 傣文语料库的构建

随着少数民族语言信息化的发展，少数民族信息化已由字处理向语言处理转变，我国少数民族语言中，蒙、藏、维吾尔、朝鲜以及彝语等有传统文字的民族都建立了语料库。2009 年 10 月，西双版纳傣族自治州建立了第一个西双版纳傣文网站，该网站是一个多语网站，包括新傣文、老傣文、汉文、英文和泰文。主要包括十二个方面内容：贝叶文化、傣乡资讯、东盟之窗、科技动态、旅游资讯、民族宗教、农业之窗、社会论坛、社会新闻、生活常识、文化生活、重要新闻。基本涵盖了傣族社会生活的方方面面。其中贝叶文化、傣乡资讯、农业之窗、重要新闻、民族宗教五个板块占语料量的 70%左右。西双版纳网站的开通，使傣族语言信息化成为可能。

3.1 语料预处理

从西双版纳傣语网站下载傣文语料，语料库中记录了傣语文本的 ID 编号、网址、所属栏目、标题、发布时间、作者、内容、出处、版面等信息。从 2009 年 9 月到 2011 年 10 月的文本，总计 727923 个音节。

由于傣文是一种拼音文字，音节与音节之间以空格分隔，从网站上下载到的很多文章中，音节之间有些出现多个空格，有些没有空格，还有些构成一个音节的字符之间出现了

换行情况，有些文章中还夹杂着不少 html 标记。因此，在进行傣文分词之前，必须先对文本进行预处理。基于傣文的特点，总结了以下六条规则，对语料进行了很好的预处理。

规则一：字符 [illegible] 、[illegible] 、[illegible] 是词尾标志，如果这三个字符后面还有其它字符，则这三个字符与其它字符之间必须加入空格；

规则二：在没有字符 [illegible] 和 [illegible] 出现的情况下，遇到 [illegible] 这些字符，则这些字符与后面的字符之间也必须加空格；

规则三：字符 [illegible] 是一个音节开始的标志，如果出现这三个字符，这三个字符前面要加空格；

规则四：如果以下字符 [illegible] 中的任意两个前后出现，则这两个字符之间要加空格；

规则五： [illegible] 必须与前后字符分开单独成音节；

规则六：᧑ 、᧒ 、᧓ 、᧔ 、᧕ 、᧖ 、᧗ 、᧘ 、᧙ （“一”到“九”）与前面任何一个字符连在一起分开，但他们之间可以不分开。目前是十为单位，如᧗᧘ 表示七十八，还是放一起。

3.2 语料切分标注

经过预处理后，首先对其中约 30 万音节进行人工标注，形成一个训练语料，并从中提取了一个 8000 多词语的词表，然后使用基于词表的方法对傣文语料进行分词，利用词表进行分词的结果见表 1：

表 1.基于词表的分词结果

算法	词表词	未登录词	总切分词数
最大正向匹配	461688	539	462227
最大逆向匹配	461317	467	461784

然后对全部切分结果进行人工校对（727923 个音节），得到 1744 篇质量较高的训练语料和测试语料，供给 CRF 模型，最终开发了一个基于音节标注的傣文分词系统，使用 CRF 进行分词，只需要少量的标注语料，系统在整体性能上就能达到 F 值为 95.58 的较高水平。实验结果表明，使用 CRF 模型，结合傣文的一些固有特点对傣文进行分词切实可行。

经过校对后的训练和测试语料格式如下：

<[illegible] #q#mu2#群#dw> <[illegible] #v#phia2phaai1#传播，散发> <[illegible] #n# kaan1#事情# bl> <[illegible] #i# iau2loa5phoa2toang5 # 旅游观光# dw> <[illegible] # ns # sip7soang1pan4naa4 #西双版纳# dwhg> <[illegible] #v#pai1#走，去#dw> <[illegible] #p#thuang1#到，到达#dw> <[illegible] #r#ti5nai1#哪里#dw>。

符号“<”与“>”之间表示一个词，每个词给出了傣文形式、词性、音标、中文解释、来源（来源指明该傣语词是汉语借词、巴利语借词还是傣文词汇等）这几个信息，各个信息之间以“#”号分隔。

4 傣文分词系统的设计

4.1 傣文分词思想

与中文类似，可将傣文分词问题看成给傣文中每个音节标注位置信息的过程，其核心本质是给傣文文本中的每个音节标注位置信息，然后通过位置信息将音节串转换成对应的合法词串。

傣文词是由音节构成的，其中单音节词由一个单独的音节构成，多音节词则由两个或两个以上的音节按顺序组合构成，因此音节与词之间存在着一种特殊位置对应关系。Xue NianWen 的中文分词系统中定义的“4-tag 标记集：词首（LL）、MM（词中）、RR（词尾）、LR（单字词）”[1]，Hwee Tou Ng 2004 年提出简化标记词首（B）、词尾（E）、词中（M）、单字词（S）[2]，在傣文分词系统中，采用这种简化的位置标记来表示音节在词语中所处的不同物理位置，则只要给定一个词就能将其中的每个音节标上一个唯一的位置标记，同样地只要给定一个带位置标记的音节序列，就能将其转化为唯一的合法词串。

目前，经过人工校对的傣文语料库文本内容都是以词串的形式保存的，因此在进行音节标注模型训练前需要将傣文词串转换为带位置标记的音节串。在利用上述四个位置标记集进行语料转换前后格式如下所示，其中(a)为原始句子；(b)为人工切分校对后的词串，(c)为带位置标记信息的音节串：

(a) ဧပ္ေ တ႙ ၵ၆ ဘ႙ ၵ တ႒၆ ဂ္ဂ္ၹ ၉ ဏ႙ ၵေ ဂ္ဒθ ၉၆ ၵ ၉ ဝ၁၆

(b) <ဧပ္ေ တ႙ ၵ၆ ဘ႙ ၵ #n#phu3koan2oan1#领导#dw> <တ႒၆ ဂ္ဂ္ၹ ၉ ဏ႙ ၵေ #n#taai2piau4thoan6#代表团#hj> <ဂ္ဒθ ၉၆ ၵ ၉ #n#tseau5hau4#我州#dwhj> <ဝ၁၆ #v#vaa5#说#hg>

(c) ဧပ္ေ B တ႙ ၵ၆ M ဘ႙ ၵ E တ႒၆ B ဂ္ဂ္ၹ ၉ M ဏ႙ ၵေ E ဂ္ဒθ ၉၆ B ၵ ၉ E ဝ၁၆ S

与基于词表的方法不同，基于音节标注的分词方法，未登录词（OOV）是指在测试语料中出现而在训练语料中没有出现的词。在传统的基于词表的分词方法（如正向最大匹配 FMM）中，如果待切分的文本包含未登录词，则这个词被正确识别出的概率是很小的。但在基于音节标注的分词方法中，词是通过相邻的音节的位置标记信息在概率上求最优解后组合而成的，在这个求解过程中词表词和未登录词是没有明显区别的。如存在某个新词，该词有两个音节组成，记为 AB，这个词从来没有在训练语料中出现过，但是以音节 A 开头的词出现过多次，以音节 B 结尾的词也出现过多次，这样在计算 AB 这个未登录词音节的位置信息时就会认为“A”被标注为词首（B）且同时“B”被标注为词尾（E）的概率比较大，这样“AB”就有很大的可能被当作一个词而识别出来。

基于音节标注的分词方法由于把词表词和未登录词放在同一个框架下进行统一处理，因此在一定程度上解决了基于词表的分词方法无法解决未登录词的问题。

---

[1]Nianwen Xue,Libin Shen.”Chinese Word Segmentation as LMR Tagging”. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL’03,2003, P.176-179.

[2]Hwee Tou Ng.Chinese partofspeech tagging: One-at-a-time or all-at-once word-based or character-based.2004

4.2 傣文分词实验

4.2.1 定义特征模板

“字符的 n 元特征在基于机器学习的中文分词中出于其离散性被经常运用”[3]。傣语借用了很多汉语和巴利语词，虽然双音节词还是主流但词长总体比汉语要长一些，四音节词也比较发达，出现了一个汉语或者巴利语，经常用傣语作解释，一个双音节的巴利语后面再跟一个傣语就组成四音节了，而且这是傣文中的一种使用习惯。曾对一篇大约 1 万多音节的傣文做过人工统计，单音节大约在 22.31%，高于汉语；双音节大约在 62.86%，低于汉语；三音节和四音节比例比重较大，两者约为 14.42%，远高于汉语。基于此，本文使用 5 音节窗口特征，它们分别是 C-2、C-1、C0、C1、C2、C-2C-1C0、C-1C0C1、C0C1C2、C-1C0、C0C1，字母 C 代表一个音节，其下标 -2，-1，0，1，2，分别代表前面第二个音节、前一个音节、当前音节、后一个音节、后第二个音节。

使用 5 音节模板窗口如下：

```
# Unigram
U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-2,0]/%x[-1,0]/%x[0,0]
U06:%x[-1,0]/%x[0,0]/%x[1,0]
U07:%x[0,0]/%x[1,0]/%x[2,0]
U08:%x[-1,0]/%x[0,0]
U09:%x[0,0]/%x[1,0]
# Bigram
B
```

4.2.2 实验结果

由于经过校对的傣文文本数量较少，对 1744 篇傣文文档采用十折交叉验证的方法进行模型训练和测试，每次使用 174 个文档作为测试集，1566 个文档作为训练集来训练模型，最终取十次实验得到的平均值作为最终实验结果。

按照 SIGHAN 的评测标准，对分词结果进行评价，包括分词整体的准确率 P、召回率 R 和综合评价 F，未登录词（基于序列标注的分词方法没有词表方法中定义的未登录词概念，这里所说的未登录词是指没有在训练语料中出现的词）的准确率 P、召回率 R、综合评价 F，非未登录词（在训练语料中出现过的词）的准确率 P、召回率 R、综合评价 F。实验结果见表 2：

表 2.傣文分词试验结果

	准确率	召回率	F值
整体	93.88	94.66	94.27
未登录词	41.43	45.12	43.21
非未登录词	95.32	95.92	95.62

[3] 姜锋：基于条件随机场的中文分词研究【D】.大连：大连理工大学，2006.

4.2.3 融合傣文特征的分词结果后处理

由于傣文中未登录词所占比例较少，因此，尽管未登录词的识别能力并不理想，但对系统整体的准确率、召回率、F 值影响不太大。经过对模型分词的错误类型进行分析，分词错误主要有以下几种类型：

(1)将一个词切成多个单音节词，这是最常见的切分错误；

例如“ဘθ ၎e ဂ◌လθ ʒ ဘၣ လ၃ ၛ”这个傣文表示“血管；动脉”，组成该词的四个音节都被 CRF 单独切分开，一个词被切成了 4 个词。这四个音节都可以单用，而且活动能力比较强。这种情况只能通过该词所处的上下文环境来对切分进行处理。但像“တာ၆ ၆ၙe”表示“大学”，这些音节只能同时出现，因为是汉语借词，一分开没有任何意义，将这类词定义为绝对切分词。通过对傣语语料库的统计分析，傣语绝对切分词具有以下特点：

- 独立成词频率较高；
- 被分开频率较低；
- 被其它词包含而成为新词的频率较低；

把这类切分错误归为绝对切分词错误。

(2)歧义切分错误：当一个傣文句子或一个傣文音节串中包含歧义时，必定是因为其中的某些音节对应着多个位置标记。确定每个音节对应的正确位置标记是通过这个音节的上下文环境而决定的，只要给定一个音节的上下文信息，就能很准确的为这个音节选择一个位置标记。所以傣文分词问题最终就是根据句子中每个音节的上下文信息给这个音节选择一个位置标记。而这个任务可以通过有指导的机器学习模型来完成。

(3)包含音节数目较多的人名、地名，尤其是借用的汉语结构名。

(4)傣文书写错误词。

通过以上的错误类型分析，结合傣文的特点，提取了一个傣文常用汉语借词词表和常用傣文绝对切分词表，对待切分语料先进行预处理后再使用 CRF 模型进行分析。

绝对切分词表和傣族汉语借词词表提取：根据傣文特点，在词表中提取出来源为“汉借”的双音节及以上的词语，提取出来源为“巴利语”的三音节以上词，以及词性被标注为“成语或惯用语”的词语，再加上人工整理的 1562 个绝对切分词，构成最终的预处理词表，该词表包含了傣语中大多数常用的绝对切分词，能很好的解决绝对切分词被模型切开的问题。

预处理的过程如下：对读入的待切分傣文文档先用提取的绝对切分词表和汉语借词词表进行粗切分，例如识别出音节串“ၹ၆ ဥ◌ၡ၆ ဂ◌ဂ◌ၛ၆ ၄ၥ”是一个绝对切分词，将整个串标注为“t”，在使用 CRF 进行标注时，如果一个串已经被标为“t”，该串不需要进 CRF 标注模型，可直接为该串中的每个音节赋予一个标记。最终得到的实验结果见表 3：

表 3.加入绝对切分词表后分词结果

	准确率	召回率	F 值
整体	95.64	95.52	95.58
未登录词	57.75	64.23	60.82
非未登录词	96.71	96.32	96.51

实验结果表明，加入绝对切分词表能明显改善整个未登录词识别的准确率。

5 结语

本文首先建设了一个傣文语料库，并根据傣文的特点，构建出一个傣语词表，基于该词表，使用最大匹配算法对傣文文档进行粗切分，对粗切分的结果经过人工校对，形成质量较高的训练和测试语料，使用 CRF 模型进行分词，并结合傣文特点，加入绝对切分词表对待切分语料进行预处理，进一步提高分词系统的整体性能，最终达到十折交叉验证平均 F 值为 95.58 的结果。该方法解决了基于词表的分词方法中由于缺乏语料，未登录词识别、歧义字段消解困难的问题。实验结果表明，该分词方法对傣语切实有效，填补了傣语分词系统的空缺，同时，达到较高切分水平的系统为后续进行傣文输入法开发、傣文自动机器翻译系统开发、傣文文本信息抽取等高层次傣文信息处理的奠定了较好的基础。

参考文献

1.Nianwen Xue, Libin Shen.“Chinese Word Segmentation as LMR Tagging”. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in conjunction with ACL’03，2003， P.176-179.
2.姜锋：基于条件随机场的中文分词研究[D].大连：大连理工大学，2006.
3.宗成庆：统计自然语言处理[M].清华大学出版社 2008 年版，第 105～129 页.
4.梁南元：书面汉语自动分词系统——CDWS[J].中文信息学报,1987,(2).
5.孙茂松,肖明,邹嘉彦：基于无指导学习策略的无词表条件下的汉语自动分词[J].计算机学报,2004,(6).
6.戴红亮：傣汉《民族区域自治法》词语统计及比较分析[J].载《构建多语和谐的社会语言生活》，民族出版社，2009 年版，第 589～597 页.

[i] 本研究受到 2010 年度国家社会科学基金项目《傣汉双语语料库建设及现代傣语词汇研究》（批准号：10MZ005）资助，在此表示感谢。