文章编号: 1003-0077(2011)04-0054-03

央金藏文分词系统

史晓东[1], 卢亚军[2]

(1. 厦门大学 人工智能研究所, 福建 厦门 361005; 2. 西北民族大学 机器翻译研究所, 甘肃 兰州 730030)

摘 要: 藏文分词是藏文信息处理的一个基本步骤, 该文描述了我们将一个基于 HMM 的汉语分词系统 Segtag 移植到藏文的过程, 取得了 91% 的准确率。又在错误分析的基础上, 进行了训练词性的取舍、人名识别等处理, 进一步提高了准确率。

关键词: 藏文分词; 自然语言处理; HMM

中图分类号: TP391 文献标识码: A

A Tibetan Segmentation System—Yangjin

SHI Xiaodong[1], LU Yajun[2]

(1. Institute of Artificial Intelligence, Xiamen University, Xiamen, Fujian 361005, China;
2. Institute of Machine Translation, Northwest University for Nationalities, Lanzhou, Gansu 730030, China)

Abstract: This paper describes the porting of a Chinese segmentation system to handle Tibetan. The F-measure of the new Yangjin system is above 91% over a test corpus although the training corpus is relatively small. It also describes more processing upon error analysis which led to further improvement.

Key words: Tibetan segmentation; natural language processing; HMM

1 引言

随着少数民族语言(主要是藏、维、蒙)到汉语的机器翻译研究逐渐进入人们的视野, 相关的少数民族语言基础语法分析工具也亟待完善。藏文分词是藏语到其他语言的基础性工具。虽然研究的时间也不算短(2002 年陈玉忠[1]是较早的一篇研究), 已经有至少 10 年的历史, 但是还没有公开可用的工具。第一作者在研究汉语分词方面有丰富的经验, 从 2005 年就开发的 Segtag 汉语分词系统, 虽然没有发表相关的论文, 但是在北京大学公开的 1998 年《人民日报》一个月的语料上的准确率约为 98%。因此将其移植到藏文, 并加以公开, 是我们的一个想法。经过与第二作者密切合作, 已经成功地开发出了藏文的分词标注系统, 在一个测试集上的准确率约为 93%, 取得了较为令人满意的效果。本文描述该系统的基本算法, 并对藏文所作的特殊改进。

本文下面的内容如下: 首先综述一下国内外的相关工作, 然后介绍了央金藏文分词系统的基本结构, 然后再描述为了改进性能对藏文所作的特殊处理, 最后得出结论, 并指出了进一步的工作。

由于第一作者一点也不懂藏文, 因此本文对想开发一个未知语种(如蒙语、泰语、彝语等)的分词系统的人, 有一定的借鉴意义。

2 相关工作

陈玉忠[1]在 2002 年提出了基于格助词和接续特征的藏文分词算法。从此文中作者得出, 其实藏

收稿日期: 2011-04-17 定稿日期: 2011-05-20

基金项目: 福建省自然科学基金资助项目(2006J0043); 福建省重点科技项目(2006H0038); 国家 863 资助项目(2006AA010108); 国家社科基金重点项目(05AYY001)

作者简介: 史晓东(1966—), 男, 教授, 主要研究方向为自然语言处理; 卢亚军(1956—), 男, 教授, 主要研究方向为藏语, 语料库语言学, 藏汉翻译。

文和日语类似, 有很多格助词, 表示一定的句法语义功能。扎西加等[2] 给出了藏文分词的词类划分。Huidan Liu 等[3] 研究了藏文分词中的数字识别问题。才智杰[4] 描述了班智达藏文分词系统的设计和实现。苏峻峰[5] 描述了一个基于 HMM 的藏文分词模型。Yuan Sun 等[6] 在天之灵 2009 年也实现了一个基于格助词和接续特征的分词算法, 并做了组块切分研究。刘智文[7] 做过一个基于 CRF 的藏文分词系统。国内的藏文相关工作基本上集中在青海师大、西北民大、西藏大学等单位。

与采用机器学习为主的汉语分词相比, 目前藏文分词系统显得落后一些。在汉语方面一般都采用 HMM、ME、CRF 等模型, 很少采用相对原始的规则或最大匹配模型。

作者也用过青海师大开发的国内最早的藏文分词软件班智达, 但是该系统只支持班智达编码。

3　央金藏文分词系统介绍

HMM 模型由于其简单高效已经成为了分词系统的基准模型, 虽然 ME 或者 CRF 的准确率比 HMM 略高一些, 但是其训练却相对复杂一些, 而且当标注语料库比较小的时候, 并不见得有优势。所以我们便用 HMM 模型来做藏文分词。另外, 作者恰好早已经实现了一个基于 HMM 的汉语分词系统 Segtag, 因此便直接移植过来。

Segtag 的结构非常简单, 分词和标注一体化完成, 其结构如图 1。

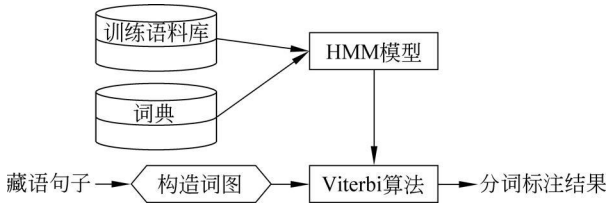

图 1　Segtag 分词系统(其中词典为央金系统所加)

由于 Segtag 本身已经是基于 Unicode 的, 所以对 Unicode 的藏文处理毫无困难, 原始程序改动不到 1%, 主要是参照文献[8] ①修改了词性表, 并增加了对藏语 Unicode 的未登录词识别。因为, 与汉语相比, 在 0 平面内, 一个汉字只需双字节表示码位, 藏文很多字(有些文章称之为字丁[9], 其实指一个可纵向叠加的书写单位, 我们仍然称为字) 是多个双字节构成的序列。此外, 专门针对藏文数字修改替换了原汉字数字识别, 使之能处理藏文数字。移植后的系统由第二作者命名为央金藏文分词系统。

如果纯粹用训练语料来生成分词词典, 由于训练语料很小, 得到的词条仅有 13 200 余条, 根本无法对藏文进行分词。所以我们又合并了几本藏文词典。大约有 9 万词。简单地把词典中的词条以频率 1 加到训练语料, 从中训练出分词词典, 一共 97 800 余条。

央金系统的性能如表 1(此处 2. 7M 指 UTF16 编码的文件大小) 。

表 1　央金分词系统的性能

训练语料	测试语料	精确率	召回率	F 值	备注
2. 7M+ 词典	25K	92. 215%	90. 041%	91. 115%	分词
		79. 342%	79. 647%	79. 494%	标注

这些训练语料都是在央金系统的分词结果的基础上, 由第二作者校对修正而滚雪球一样得到。而初始种子语料来自于班智达分词系统。

另外, 虽然 Unicode 目前已经是国际标准, 国内仍然存在着部分班智达和同元编码的文档, 我们集成了编码识别和自动转换功能, 以方便用户使用。

此外, 我们还集成了鼠标藏汉词典, 以方便作者校对分词结果。

由于第一作者一点也不懂藏文, 所以许多央金分词系统的很多功能都是为了方便用户能够在系统内便于进行分词校对而设。

4　分词系统的错误分析和改进

4. 1　分词系统错误

通过文件比较, 对测试语料中的错误进行了分析。首先我们注意到, 标注的准确率偏低。结果发现, 训练出词典中的有些词的不同词性之间的频率差异很大, 如

ར gj 1 tt 1 nn 9 vi 20 gl 3413

其中 gj 和 gl 都是格助词, 怀疑 gj 这个词性是训练语料中的标注错误而混进来的, 因此在装入词典的时候做了一个简单的处理: 如果某个词的频率低的词性与该词的频率最高的词性之频率比小于阈值 β(目前取 1%) , 则舍弃该词性。经过这样处理以后, 分词的准确率没有任何变化, 而标注的准确率有所提高。

① 实际上我们参考的主要规范是青海师范大学才让加、吉太加、扎洛等起草的拟作为教育部标准的“信息处理用藏语词类标记规范”。

简单的分析表明: 分词错误大部分是由于未登录词而造成的。而很多标注错误是因为训练生成的词典中根本没有测试答案中的词性造成的。其实这些错误大部分是训练语料的不一致性造成的。

舍弃低频词性后央金分词系统的性能见表 2。

表 **2** 舍弃低频词性以后央金分词系统的性能

训练语料	测试语料	精确率	召回率	F 值	备注
2. 7M + 词典	25K	79. 342%	79. 647%	79. 494%	原系统
		82. 632%	82. 949%	82. 790%	改进 1

4. 2 汉语人名识别

藏文新闻中经常出现人名。相对于地名等其他专名, 人名是最丰富并且变化的。因此, 分词系统最好能自动识别人名。从来源分, 人名基本上可以分为藏语人名、汉语人名、欧美人名等三大类。目前我们只考虑了汉语人名的自动识别。

汉语人名翻译成藏语, 基本上都是采用音译。也就是说,“王东”和“王栋”翻译成藏语应该是一样的。当然, 不同的译者可以选择不同的藏文字来对同一个汉字(或同音汉字) 进行译音。目前我们已经收集了一个汉藏人名对照表 TC(目前只有 300 条) , 我们可以把它改为藏音对照表(这里音指汉语拼音) 。另外我们还有一个常用汉语人名表 C, 有 20 多万条。此外还有一个海量的汉语语料库。那么藏文中的汉语人名识别算法可简单地描述如下:

假设藏文的音节序列 ABC, 其中每个音节都是一个可能的汉字译音 A′B′C′, 而且不是藏文单词, P(A′B′C′) 作为汉语人名的概率大于一定的阈值, 那么可把 ABC 识别为一个藏文中的汉字人名译音。

人名识别和数字识别都在图 1 的构造词图中进行, 与其他处理无关。其实实现的时候就是和数字识别一样加一个加权自动机即可。

人名识别后的央金分词系统的性能见表 3。

表 **3** 人名识别后的央金分词系统的性能

训练语料	测试语料	精确率	召回率	F 值	备注
2. 7M + 词典	25K	92. 119%	92. 473%	92. 296%	分词
		83. 015%	83. 333%	83. 174%	改进 2

尽管有所改进, 但和汉语分词相比差距不小, 训练语料库太小可能是一个主要原因。

5 结论和进一步的工作

本文描述了一个基于 HMM 的藏文分词系统。就我们和同类系统比较而言, 该系统的分词速度快, 准确率也基本达到了可以使用的水平, 目前已经用于我们的藏汉统计机器翻译系统。

下一步要做的主要工作是: 继续扩大训练语料规模; 进行地名和机构名的自动识别; 克服 n 元模型的局部性, 处理长距离语义相关性。

参考文献

[1] 陈玉忠, 李保利, 俞士汶. 藏文自动分词系统的设计与实现[J]. 中文信息学报, 2003, 17(3) : 15-20.

[2] 扎西加, 珠杰. 面向信息处理的藏文分词规范研究[J]. 中文信息学报, 2009, 23(4) : 113-117.

[3] Haidian Liu. Tibetan Number Identification Based on Classification of Number Components in Tibetan Word Segmentation[C] // Proceedings of the Coling 2010: 719-724.

[4] 才智杰. 班智达藏文自动分词系统的设计与实现[J]. 青海师范大学民族师范学院学报, 2010, 12(2) : 75-77.

[5] 苏峻峰. 祁坤钰, 本太. 基于 HMM 的藏语语料库词性自动标注研究[J]. 西北民族大学学报(自然科学版) , 2009, 30(1) : 42-45.

[6] Yuan Sun et al. Design of a Tibetan Automatic Word Segmentation Scheme[C] // Proceedings of International Conference on Information Engineering and Computer Science, 2009: 1-6.

[7] 刘智文. 藏汉统计机器翻译研究[D] . 厦门大学硕士论文, 2010.

[8] 才让加. 藏语语料库词语分类体系及标记集研究[J]. 中文信息学报, 2009, 23(4) : 107-112.

[9] 王维兰, 陈万军. 藏文字丁、音节频度及其信息熵[J]. 术语标准化与信息技术, 2004, (2) : 27-31.