彝文网页文本分词平台①

孙善通, 王嘉梅, 李炳泽, 胡 刚

(云南民族大学 电气信息工程学院, 昆明 650500)

摘 要: 在机器翻译、自动分类、搜索引擎等技术中, 彝文分词具有很重要的作用, 同时也是彝文信息处理至关重要的环节. 本文以当前的彝文分词技术为基础, 通过构建彝文词库, 并用彝文网页获取平台抓取彝文网页文本, 结合彝文特有的优势, 从分词词库、分词算法、结构流程、系统界面和模块、实验结果等方面进行了详细的分析, 最终实现彝文网页文本分词平台. 最后的结果表明, 本平台分词准确率较高, 实用性和通用性也较好.

关键词: 彝文网页; 词典分词; 词库; 彝文分词; 分词平台

Analysis and Discussion of Yi Word Segmentation

SUN Shan-Tong, WANG Jia-Mei, LI Bing-Ze, HU Gang

(School of Electrical and Information Technology, Yunnan Minzu University, Kunming 650500, China)

Abstract: In the fields of machine translation, automatic classificationand search engine technology, Yi word segmentation plays a very important role, which is also a vital part in Yi language information processing. This paper is based on the current segmentation of Yi word.Through the construction of Yi thesaurus and webpages of Yi, we can grab the page texts of Yi. Combining with the advantages of Yi language, with a detailed analysis such as the thesaurus, word segmentation algorithm, flowchart and structure, system interface and modulesand the experimental results, we build the segmentation platform of Yi page text. Finally, the results show that the segmentation platform has a property of higher accuracy, practicality and versatility.

Key words: Yi web pages; segmentation dictionary; thesaurus; Yi word segmentation; word segmentation platform

进入 21 世纪以来, 信息技术越来越发达, 越来越多的科研人员投入到彝文的信息处理工作中, 使得彝文信息的处理发展迅速[1]. 但要达到自动化的目标, 语言处理阶段就是必经环节. 计算机文字信息处理发展至今, 汉语分词研究最为完善, 其他主要少数民族语言分词研究, 例如藏文、泰文、蒙文等发展也较好, 而彝文分词的研究相对滞后[2]. 彝文自动分词对于将来的彝语信息处理中的文字识别的自动化、文本的校对自动化、词汇获取自动化以及实现彝文的翻译自动化等等都是很好的铺垫, 同时也为以后的彝文信息处理奠定了良好的基础.

处理彝文自然语言时, 彝文词是以基本操作单位以及重要的信息载体出现的, 不仅在自然语言信息处理中还包括人工智能应用研究中, 彝文分词都很重要, 具有不可替代的实用价值[3]. 众所周知, 计算机语言中的彝文文字信息处理所涉及到的信息检索、语义分析、语法分析再加上机器翻译等等, 这些方面应用到的基本语言信息处理单位都是以彝文词为基础, 但是彝文词与词之间的切分标志显得不那么明显, 这样给彝文信息检索的开展、机器翻译、语义分析等方面的处理增加了一定的难度, 这就要求彝文信息处理之前关键是解决彝文中词与词之间的切分问题[4]. 本文以对彝文语言文字分析为基础, 具体地由彝文语料库设计、算法、分词规则、性能测试、结构流程等方面对彝文网页文本分词平台的设计与实现进行了探讨, 对彝文网页文本分词平台的研究进行尝试性探索.

① 基金项目:国家自然科学基金(61363085)

收稿时间:2016-02-24;收到修改稿时间:2016-03-28 [doi: 10.15888/j.cnki.csa.005399]

1 彝文词库设计

从某种意义上来说，彝语与汉语具有一定的相似之处，比如说字的大小，两者基本相同，汉语中的“方块字”，在彝语中则称为“石块字”[5]．此外，都是以字为分界，词与词之间的界限都不是特别明显，彝语与汉语分写时都不是把词作为基准[6]．与此同时，彝语词不同于汉语，彝语词的分词标志很模糊，没有显而易见的词头和词尾等标志．

本文在相关彝文专家研究的基础上，借鉴前人彝文分词研究经验，并基于彝文网页词语的特点，查阅相关彝文资料，最终确定了彝文词条来源，同时制定彝文词库设计原则[7]．收集的词条主要来源于《彝汉四音格词典》、《彝汉字典》、《彝语大词典》、《彝族比尔词典》、《滇南彝文字典》、《彝文字集》、《彝文字典》，词库最终收集了约 8 万多条彝文词．词库有教育科学、数学物理、生理卫生等 17 个分类，除此之外还包含相对应的中文、英文、词性和拼音信息，如图 1 所示．

	A	B	C	D	E
1	中文	彝文	简称	拼音	英文
2	安培	[illegible]	v	ān péi	Ampere
3	弹性碰撞	[illegible]	n	tán xìng pèng zhuàng	Elastic collision
4	等号	[illegible]	n.	děng hào	Equal sign
5	电荷	[illegible]	n.	diàn hè	Electric charge
6	电荷守恒定律	[illegible]	n.	diàn hè shou hén	The law of conservation of charge
7	电荷数	[illegible]	n.	diàn hè shù	Charge number
8	非弹性碰撞	[illegible]	n	fēi tán xìng pèng zhuàng	Inelastic collision
9	分毫	[illegible]	n.	fēn háo	Nothing
10	风化	[illegible]	v.	fēng huà	Weathering
11	负电荷	[illegible]	n.	fù diàn hè	Negative charge
12	负号	[illegible]	n.	fù hào	Minus
13	感生电荷	[illegible]	n.	gǎn shēng diàn h	Induced charge
14	合数	[illegible]	n.	hé shù	Composite number
15	恒等式	[illegible]	n.	héng děng shì	Identity
16	恒定电流	[illegible]	n.	héng dìng diàn l	Constant current
17	宏观物理现象	[illegible]	n.	hóng guān wù lǐ x	Macroscopic physical phenomena

图 1 彝文词库结构图

2 彝文分词算法

如今相对主流的分词算法主要有如下三种：第一种是基于词典分词；第二种是基于统计分词；第三种是基于理解分词[8]．基于理解的分词算法实现相对困难，以下仅对基于词典分词分词算法和基于统计的分词算法做详细介绍．

2.1 基于词典的分词算法

基于词典匹配的分词算法(又称为“字符串匹配分词算法”或“机械匹配分词算法”等)，相应的词典是许多适用的分词系统的必备工具，根据提供的词典进行匹配来识别每句话中出现的词语．使用词典匹配的分词算法有多种，其中包括正向最大匹配、反向最大匹配、全切分、最少词切分等．对于词典匹配分词算法，我们以正向最大匹配分词算法为例：

正向最大匹配分词(FMM)是最简单的一种分词算法．该算法首先构建词典，并设计基于该词典的第二种匹配算法．然后按照如下过程进行分词：

第一，假设 S 为准备切分的语句，W 为词串切分的最终结果；

第二，假设 i 为将要切分的位置，对 i 进行初始化设置，设为 1，且 $i \le |S|$；

第三，将字串 S 中切分出的最长词 w 放在词串 W 之后，位置 i 的值变为 $i + |w|$；

第四，返回到第三步，当位置 i 变到字串 S 结尾时终止．

2.2 基于统计的分词算法

分词也可以看作是一个对应问题，给定一个彝文序列，找对应的词序列．设待切分的彝文字串为 $S = s_1 s_2, ..., s_n$，彝文字串 S 对应的某一种切分方式为彝文词串 $W = w_1 w_2, ..., w_n$．切分方式 W 的概率：

$$P(W \mid Y) = \frac{P(Y \mid W) * P(W)}{P(Y)}$$

如上文所描述的，全切分就是将句子的每一种切分方式罗列出来，根据统计学的方法求出最大可能性的切分方式，即求出一个最大的 $P(W|Y)$．条件概率 $P(Y|W)$ 对于每一种切分方式都一样，均为 1．而 $P(Y)$ 是每一种切分概率计算的共同项，可以去掉．因此，可以直接求 $P(W)$．对于统计分词算法，我们以最大概率分词算法为例：

假设 W 为某一词串，且 W 中的词与词是相互独立互不相关的，则在使用最大概率分词算法求解 $P(W)$ 的计算公式如下：

$$P(W) = \prod_{i=1}^{m} P(w_i) = P(w_1) * P(w_2) * ... * P(w_m)$$

当使用最大概率分词算法时，分词的最终结果是以概率最大的为准的．对于具体的某个词的概率计算，则使用如下公式：

$$P(w_i) = \frac{\text{语料库中 } w_i \text{出现的次数}}{\text{语料库总词数}}$$

彝文分词基本的原则是切分出的彝文词在最大程度上保持长度，使得切分出的彝文词的数量减到最少．本文根据彝文网页文本的特点，采用基于词典与统计相结合的彝文分词法．

3 彝文网页文本分词平台实现

由彝文网页文本分词平台软件大小和实现切分目标来看，彝文网页文本分词平台运行所需的硬件条件要求并不高，在普通的笔记本和台式机上都可以流畅运行，操作系统选择的是微软的 Windows7. 结合本平台相关功能模块的实现需求，软件开发时选用 C++进行开发设计，并利用 Microsoft Visual Studio 2010 平台来实现彝文分词设计开发.

3.1 核心代码

彝文网页文本分词平台结构如图 2 所示.

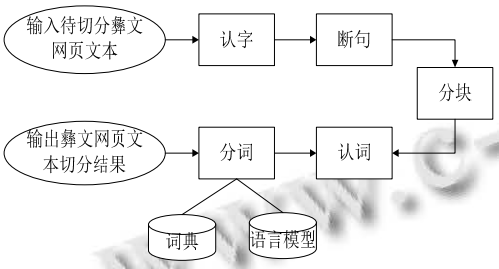

图 2 彝文网页文本分词平台结构图

彝文网页文本分词平台主要包括 5 个部分：认字、认词、分词、断句和词库维护. 第一认字：彝文网页文本分词平台对所获取的彝文网页文本进行彝文字体识别；第二认词：彝文网页文本分词平台根据彝文词库对获取的彝文文本进行认词；第三分词：彝文网页文本分词平台对获取的彝文文本进行分词并输出分词结果；第四断句：彝文网页文本分词平台对获取的彝文文本进行断句处理；第五词库维护：词库维护人员通过添加、删除等手段对彝文词库进行管理及维护，且提供词库的导入、导出功能.

彝文网页文本分词核心代码如下:

```
if (出现过的词.Count > 0)
for (int kk = 0; kk < 出现过的词.Count; kk++)
    {if ( 出现过的词[kk].ToString().Contains(dt.Rows[jj][" 彝文"].ToString().Trim()) || dt.Rows[jj][" 彝文"].ToString().Trim().Contains( 出现过的词[kk].ToString()))
    是否出现 = true;}
    if (!是否出现)
    { 出现过的词.Add(dt.Rows[jj][" 彝文"].ToString().Trim());
a.content = a.content.Replace(dt.Rows[jj][" 彝文"].ToString(), " " + dt.Rows[jj]["彝文"].ToString() + "/" + dt.Rows[jj]["简称"].ToString().Trim());
                }
    }
最大值 = 1;
}
    if (最大值 == 1)
     出现过的表.Add(dt.TableName);
    }
     }
     DataRow dr = dtt.Rows.Add();
     dr["id"] = 当前 id;
     this.Invoke((EventHandler)delegate
            {
             dr["标题"] = drr["title"];
               });
      Random r = new Random();
   int ttt = r.Next(出现过的表.Count() - 1);
```

3.2 彝文网页文本分词平台模块和功能

彝文网页文本分词平台采用 C++进行编程实现，基础类库选用微软的 MFC 方式来设计实现. 采用此方法编程实现的彝文网页文本分词平台不仅可以保证源代码的稳定，还可以保证彝文分词的高效性，完全遵循面向对象型的程序设计思想. 彝文网页文本分词平台主要包括网页采集、词库管理和文本分词三个模块.

(1) 网页采集模块

网页采集可以获取彝文网页中文章的标题、发布时间、来源及其对应的 URL，并且通过对标题的搜索可以快速找到想要查询的信息，双击每一行都可以显示该行文章的详情信息. 图 3 为网页信息采集界面.

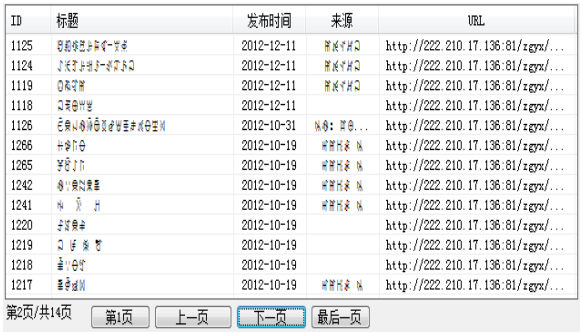

图 3 网页采集模块图

(2) 词库管理模块

词库管理模块可以通过中文、彝文关键词或者类别词汇进行检索，同时也可以导入新的词条来对词库进行更新管理，如图 4 所示.

图 4　词库管理模块图

(3) 文本分词模块

文本分词模块对采集到的彝文网页文本进行分词，框图中间部分为采集到的彝文网页文本内容，下面为对应的分词后的文本，词与词之间用反斜杠“/”分开，并在其后注明分词后的词性. 通过对一段彝文文本的切分测试，效果如图 5 所示.

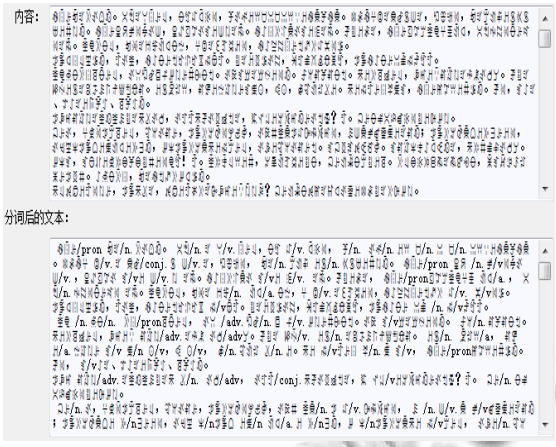

图 5　文本分词效果图

3.3 测试结果

因为彝语没有和汉语一样标准的分词参考，所以在对本彝文网页文本分词平台进行测试时是以人工分词结果为标准来进行的. 为了保证测试的科学性，选取了不同种类的网页文本进行测试，准确率达到了 90%以上，切分速度也较快，测试达到了预期的效果. 与西南民族大学开发的只采用词典分词算法的彝文分词系统相比较，本平台由于采用的是基于词典和统计的彝文分词算法，虽然在切分速度上相对较慢，但分词准确率有了一定的提升.

此外，因为彝文网页内容涉及的面较广，说明本彝文网页文本分词平台能够较好地切分不同的彝文信息，这也在一定程度上系统的通用性很强. 而且本平台界面简单，易于安装、更新和操作，可扩展性和移植都较好. 但是也存在一定的不足，没有对专有名词的识别，同时歧义字段的处理也有待于进一步提高和完善.

4　结语

本文根据彝语语素、词汇、语法、语义等方而的特点和规律，给出了彝文词库的设计思路和彝文网页文本分词规则，在分析彝文网页文本特点的基础上选用了 C++作为本平台的开发设计语言，并在 Microsoft Visual Studio 2010 平台上编程实现. 理论上，彝文网页分词平台采用词典分词法，算法采用了正向最大匹配算法进行匹配分词；实现上，给出了彝文网页文本分词平台的功能模块、主要算法和实现流程. 最后对彝文网页文本分词平台进行了测试和分析，测试结果表明，本平台分词的速度较快，准确率也较高，但对于词库中未登录词的识别能力较差，还需在后续的研究中不断改进. 本文为彝文网页信息处理在词性标注、信息检索、篇章理解、机器翻译、语义分析、语音合成等方而的应用与开发做了基础性的工作，同时也为彝文网页资源的整理和彝文网页信息的挖掘提供了一种新的方法和工具.

参考文献

1 沙马拉毅.计算机彝文信息处理.凉山大学学报,2001(12):4–8.
2 沙友富.彝文分词系统的实现及性能评价.今日湖北旬刊,2014.
3 王成平.彝文信息处理自动分词技术的研究现状与难点分析.电脑知识与技术,2012,(4):944–946.
4 金骋.信息处理用彝文特殊词类分词规则概述.大观周刊,2013,(10):3.
5 林春,王嘉梅,张建营,等.彝文网页敏感信息监测系统的研究.中国新通信,2013,(24).
6 陈顺强,沙马拉毅.信息处理用彝文词性研究.民族语文,2011,(4):49–51.
7 刘赛,李益东.彝文文字识别中的文字切分算法设计与实现.中南民族大学学报(自然科学版),2007,26(3):70–72.
8 陈顺强,马嘿玛伙.基于隐马尔科夫模型的彝文分词系统设计与开发.西南民族大学学报(自然科学版),2012(1):146–149.