多少数民族语言语音-文字处理集成框架

Multilingual Acoustic-Character Processing Integrated Framework

多种少数民族语言文字及语音语料库的建立及维护是一项需要巨大人力和精力的工作。由于，规范不统一、缺乏及时更新、缺乏人力物力等原因，很难获得高质量语料库。而且重复工作过多，缺乏交流，缺乏统一平台等导致这项工作很难前进。因此，我们开始精炼这部分工作，做出一个统一框架。希望大家共同努力，完善这项工作。

本框架有三个层次构成：（1）最底层是多语言音素-文字处理类 Mchar（）；完成各种文字的归一化，音节分析等任务。（2）中间层是各个语言特定的语音-文字处理部分，每个语言有一个类 M2chars; 完成一些语言独特的操作，如：音节规则，发音辞典，特殊发音辞典等。（3）上层给用户提供输入输出服务；如：针对某一个文本文件的文本归一化，各种粒度单元的统计及输出，发音辞典的建立等。该层通过一个对象 charObj 来完成独立于具体语言的操作。图 1 显示总体框架及主要变量和属性。

(1) 基类 Class：Mchar（）， 主要包括 4 个功能，请尽量调用这些功能不要在子类中覆盖，除非不得已。若出现错误，请通知我，多谢。（1）Code_init（）方法：从文件 XXX.code 文件中调出基本字符代码集合以及所对应的归一化 ASCII 代码。（2）Code_flip（）方法：将文本文件中的各种字符代码（如：unicode）转换成规定的 ASCII 代码文件。 （3）Syll_split()方法：能够切分维、哈、柯、汉等多种语言的音节切分任务，并输出各种切分结果，其中第一种切分正确的概率很大。各个语言只提供其音节模板即可。请尽可能调用该方法。（4）字符判别，元音判别等通用功能，各个语言只提供字符代码范围及元音列表即可。

（2）各个语言类 Class:XXXchar（），继承了 Mchar（）基类主要功能是：拼写检查及发音词典的生成。请尽量缩短该部分程序，尽量调用基类 Mchar（）。有些语言（如：Uyghur）其拼写规则，音节规则等较简洁，发音词典容易生成。有些语言需要较多预处理。

（3）文件处理类：class：Mcount（），该类没有继承其他类，是个单独类：主要提供用户接口。通过生成对象 charObj 来调用 Mchar（）和 XXXchar（）类，出来预留的“_ -”等词素标志或复合词标志以外 和语言无关，请一定保持该类与语言无关。主要功能有：（1）某一个 unicode 文件或词的代码归一化，及统计字符，音节，词等单元的内部方法。（2）Line_process()方法是个 generative 函数，实时输出单词及其他缩写词等单元。其他方法尽量调用该方法。（3）Token_Vocab()方法：收集某个文件中的字符，音节，词，未知字符，缩写词等。然后交给 Particle_export()方法来输出这些单元信息。（4）File_export()方法，以文件形式输出各种单元，如发音文件建立，发音词典建立，音节文件及音节词典建立等。除此之外，提供一些 recipe（） 方法来讲解了使用方法。

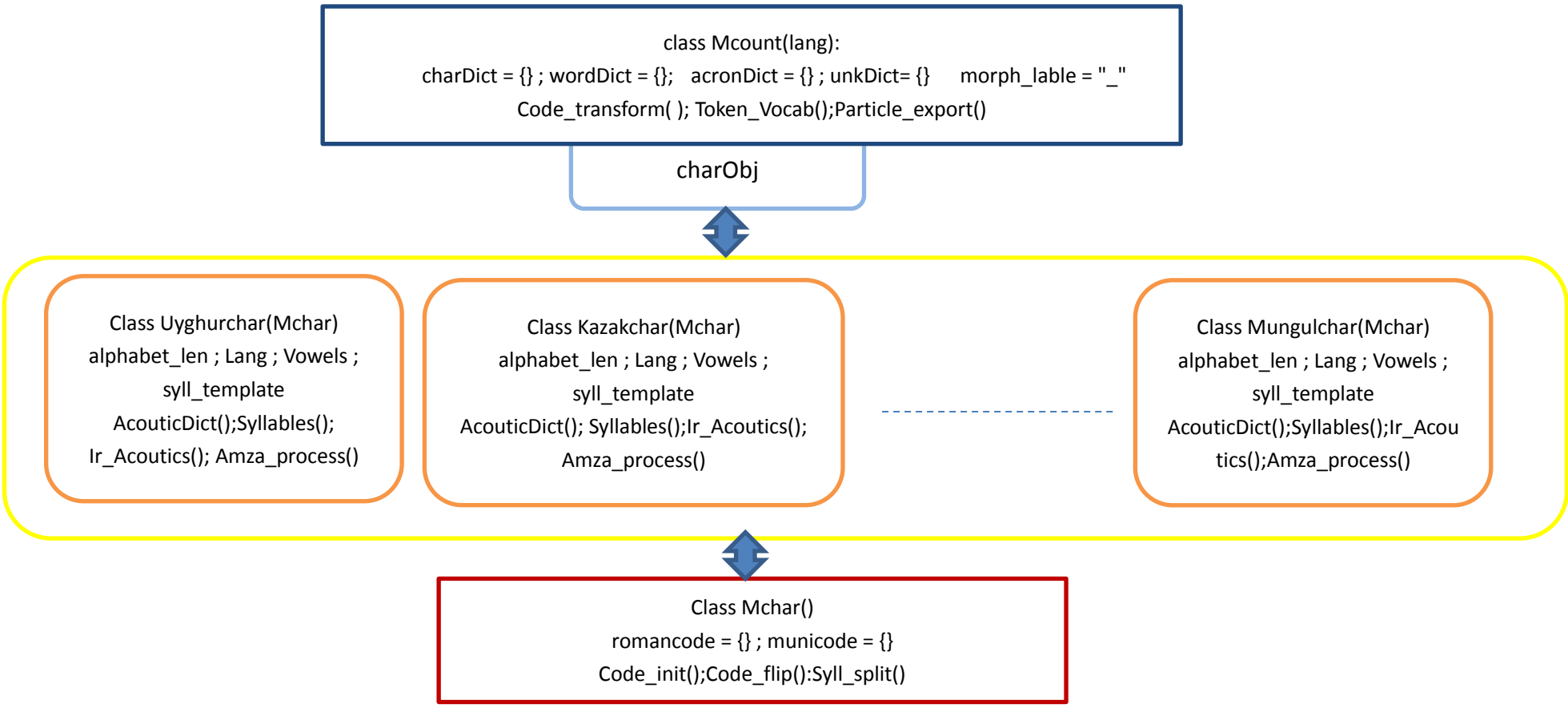

图 1 多少数民族语言语音-文字处理集成框架图