

# 云计算技术在中介语口语语料库建设中的应用<sup>①</sup>

林君峰<sup>②</sup>

## The Application of Cloud Computing in the Construction of Inter-language Spoken Corpus

Lin Junfeng

[Abstract] Today's large-scale construction of foreign language spoken corpus is

## 1 引言

二语习得理论、中介语理论建设正在如火如荼地开展,其建设成效的显现与二语习得理论建设所应更紧密地结合。“语料”可以考察学习者语言中词汇、语法、语式、语用等方面的实际表现之外,还可以了解学习者实际的口语语音面貌,可以对其进行声、韵、调等方面的考察与分析。林君峰,福建师范大学,兴福天学等院校都有汉语中介语语料库,预计建成规模最大的“全球汉语中介语语料库”(目标5 000万字)和收录教育共语三千小时时长

<sup>①</sup> 本成果受教育部哲学社会科学专项重点攻关项目“全球汉语中介语语料库建设”项目(项目编号:12JZD018)。

<sup>②</sup> 作者简介:林君峰,福建师范大学海外教育学院讲师,研究方向为语料库建设与应用、计算机辅助教学。

的口语语料。

然而,在口语语料库建设规模不断扩大的同时,其建设方式仍主要依靠人工。口语音频语料的人工转写工作相比书面语语料费时费力,转写过程中经常需要反复播放,一份音频文

字时,需要占用的服务器运算资源、带宽等也比较大,费用成本会高一些,这也可能会影响建设单位对外开放语料库的意愿。

因此,有必要探索建设语料库的新模式,降低建设人员负担,降低运营成本,提高建设效率和建设质量,并促进语料库的开放共享。

## 1.2 云计算技术的优势

云计算技术是“通过互联网提供各种计算服务和存储服务”的新兴技术,“云服务供应商”

对于用户来说,云计算技术将原本自主掌控的服务器承载的数据存储、计算等业务,分散到自主云计算系统管理的服务器上,用户不再需要直接维护服务器硬件,软件性能和安全有保障。云计算环境下,服务器运营成本显著下降,维护工作得到极大简化,同时性能更加稳定,数据安全也更有保障。

在国内互联网上,只有云、W、服务器、云存储、云语言等各种公有云计算服务可供使用,这些云服务都可以整合起来用于汉语中介语口语语料库的建设。云服务器可用于运行语料库建设及管理系统,云管理可用于语料转写,云存储可用于音频文件存储、加工和检索。

## 2 语音转写实践

口语语料库建设流程整理大致可分为采集语料资源整理、音频转写及文本、语音标注及声调标注、建立数据库、转写检索程序等环节,其中转写和标注的人工耗时最长。书面语语料的标注已能借助语料标注软件或人工标注机完成标注工作(陈建林,2019)。口语语料的标注也类似。但以往汉语语料标注技术已有了很大的发展与进步,口语音频材料的转写如

何,目前“语音”标注的准确率已接近书面语标注准确率,甚至接近书面语标注准确率(陈建林,2019)。此外,随着语音识别技术的发展,语音识别准确率已接近书面语标注准确率,甚至接近书面语标注准确率(陈建林,2019)。因此,在语音标注方面,目前仍存在一定的难度。不过,二语学习者;语音标注的语音面貌,二语水平参差不齐,国内的语音平台对二语学习者提供的标注效果如何,能否达到实用化的水平,还有待进行实际的尝试。

本文整理了中外语音识别及标注软件,标注音文件,并分门别类整理,使用自拟程序连接数据库开放平台<sup>①</sup>,进行了批量的自动标注实践。

① <http://xyyin.haidw.com>

## 2.1 音频材料的准备

将录音转换为百度语音云服务平台所支持的格式,再从中选取一部分录音剪辑另存,生成一批单个录音时长不超过 60 秒的音频文件<sup>①</sup>。

### 2.1.1 选取测试录音

在中级、高级阶段各选取了一位学生的 HSKK 模拟考试录音,一位是蒙古学生(口语一般,参加中级口语模拟考试),一位是越南学生(口语较好,参加高级口语模拟考试)。

.....

..... 的录音,考生回答部分的空白录音,9. 考场“铃”,考试结束铃声,考场外“铃”之后的录音等。

..... 截后,转换音频格式。截选完成后保存录音文件为 Windows PCM 格式文件(后缀名为.....

..... wav),作为原始录音文件,再将原始录音文件另存一份,音频采样频率调整为百度语音云

..... 服务平台所支持的 16 000 采样率,16 位单声道,以便于下一步剪辑。

### 2.1.3 剪辑测试样本

从已转换格式的 2 份录音文件中各剪辑出 12 份小文件,内容分别为:

- (1) 对姓名的提问和回答;
- (2) 对国别的提问和回答;
- (3) 对考生序号的提问和回答;
- (4) 引导语,提醒考生接下来要跟读句子;
- (5) 引导语说明要求,朗读,考生朗读 1 个句子

..... (B) 引导语说明,考生跟读 3 个句子;

..... (C) 引导语朗读,考生跟读 3 个句子(每个句子的朗读时长不超过 10 秒);

<sup>①</sup> 百度语音云服务平台当前只能上传不超过 60 秒的音频文件,超出时长则报错,不能识别。

(1) 完全正确,人工重播录音后,可直接确认通过的。

(2) 虽有少量错误但仍成句,在重播录音后可在已识别出的前后文基础上快速订正的。

(3) 部分识别,但不成句,需重播录音并听较多内容听的。

(4) 基本不成句,大部分内容必须重新录音由人工转录的。

例 5: 兴趣系指好了两声。(人工转录:兴趣是最好的老师。)

### 2.2.2 识别准确度统计

按照上述标准,对录音样本中二语者发音部分的识别准确度进行了统计,统计结果如下:

2	交替发音(共 2 个短句,二语者 1 句)	A	B
3	交替发音(共 2 个短句,二语者 1 句)	D	A
4	只有母语者发音(3 个句子)	—	—
5	交替发音(共 2 个句子,二语者 1 句)	C	A
6	交替发音(共 3 个句子,二语者 1 句)	C	A

注:表中 A 表示完全正确, B 表示有少量错误但仍成句, C 表示部分识别, D 表示基本不成句。

8	交替发音(共 6 个句子,二语者 3 句)	DAD	ABA
9	交替发音(共 6 个句子,二语者 3 句)	CCD	ABA

① 这里忽略标点的差异,因为语音识别得到的文本基本上全用逗号。

(续表)

录音样本	样本说明	蒙古学生 (HSKK 中级)	越南学生 (HSKK 高级)
10	只有母语者发音(3个句子)	—	—
11	只有二语者发音(3个句子)	30	30
12	只有二语者发音(成段表达,50秒左右,若干个句子,包括长句)	CCDD	BBBBBBBB
统计	句数(二语者发音)	共 19 句	共 23 句
	A	15.8%(3句)	34.8%(8句)
	B	10.5%(2句)	60.9%(14句)
	C	31.6%(6句)	4.3%(1句)
	D	42.1%(8句)	0%(0句)

### 2.3 可行性分析

从表 2 可以看出,母语者发音(3 个句子)的录音样本,在全部样本中只占 1.4%,占 19.4%,而大部分句子(本语群人工标注)句型简单,词汇从部分词库中选取,在语速、音量等方面又快或录入,导致其中很多部分的音频,因为考虑的二语水平一般,所以识别效果不佳,仅 C 级及以上的比例为 57.9%。

从识别结果,可以看出,对于母语者发音(3 个句子)的录音样本,语音识别准确率很高,语音识别可以完成大部分的转录工作。即使是二语水平一般的音频,自动语音识别也能减少相当一部分的人工标注量。因此,在语料库建设中,使用语音识别软件进行语音标注,对于语音识别的标注在与人工标注结果对照时,可以帮助发现一些错误。

例 6:你老板的办公室在对面。(人工转录,李经理的办公室在对面。)

例 7:对不起,这里严禁抽烟。(人工转录,对不起,这里禁止抽烟。)

例 8:如果老板不下面我们就走了。(人工转录,如果老板不下面我们就走了。)

例 9:可是两个人在一切物件打电话的时候。(可是两个人在打电话的时候。)

① 全部样本识别结果和人工听录文本下载:<http://www.banyu123.cn/html/yuliao/>

通过云语音批量自动识别音频语料,辅助口语语料转写工作,显著提高工作效率,减轻人工工作量并对错误标注有所帮助。

### 3.2 开发及运行环境配置

基础文件格式:以 XML 为基础文件格式,用于保存音频文件信息、语料转写文本等。

“XML……是 Web 服务领域的‘世界语言’,‘通用且容易解析的 XML 将会成为主流的数据交换格式。’”(单东林、张晓菲、魏然,2012:183)使用 XML 有利于程序和数据维护,便于以后对话料的检索和数据挖掘。

开发环境:编程工具为微软 Visual Studio 2013 集成开发环境(编程语言为 C#4.5),数据库使用能很好地支持 XML 的 SQL Server 2008。

七牛云存储<sup>①</sup>,云语音开放平台为百度语音。七牛云存储可免费存储 10GB 的文件,每月提供 10GB 免费流量,百度语音则完全免费。

预加工好的音频文件存放在云存储平台,相关文件信息存放在 Web 服务器的数据库中。自动转写时,Web 服务器同时连接到云存储和云语音开放平台,从数据库中批量读取识别结果保存到数据库中。

这样的配置,可以使自动转写过程中的数据传输都在云服务器间进行,运行更加稳定。

### 3.3 转写流程

#### 3.3.1 音频预加工

使用 Adobe Audition 音频编辑软件对原始音频内容做无损的加工,仅删除较长时间的空白或纯噪音的部分,不改变音频格式、采样率。(为便于叙说,将加工后的音频文件称为文件 A)

将文件 A 转换为语音云平台指定的音频格式,并拆分出若干个小文件(文件系列 B)。拆分出的单个文件,其时长不能超过云语音平台的限制(百度语音以 60 秒为限)。拆分时记录相对原文件(文件 A)的时间起止。拆分并不影响文件 A 本身。

#### 3.3.2 上传音频文件

将文件系列 B 通过 Web 服务器上传至云存储,同时将相关文件信息保存在 Web 服务

<sup>①</sup> <http://www.qiniu.com/>.

### 3.3.3 批量语音识别

在 Web 服务器选定要识别的文件范围(文件系列 B),并从数据库中批量读取文件信息。根据文件信息,调用云存储,将存储在其中的音频文件(文件系列 B)提交到云语音平台进行批量识别,并将返回的识别结果文本以 XML 格式保存到数据库中。

### 3.3.4 人工核对

人工听录音(文件系列 B)并校对识别结果文本,完成语音转写为文本的工作,并可顺带进行部分偏误标注。

### 3.3.5 合成语篇文本

在人工听录音并校对识别结果文本的基础上,根据语料库中已有的语料,按照一定的规则,将语料库中的语料进行合成,生成新的语篇文本。语料库的开放共享也有一定的作用。

## 4 小 结

不断扩大的语料库建设规模,与仍主要依靠人工的建设方式是一对矛盾。提升语料库建设的计算机技术含量有助于减轻建设人员负担,提高建设效率和建设质量。

经本文小规模测试,在口语语料库建设中应用云计算技术可以起到较好的效果,能节省

[1] 张宝林,崔希亮.谈汉语中介语语料库的建设标准[J].语言文字应用,2015(2).

[2] 朱文武.多媒体云计算[J].电子产品世界,2011(9).

[3] 百度云存储开放平台. <http://www.baidu.com/docs/asi/>.

[8] 七牛云存储音视频/流媒体在线处理. <http://www.qiniu.com/feature#data-process>.

[9] 讯飞语音云开放平台. <http://www.xfyun.cn/index.php/services/voicedictation>.