云计算技术在中介语口语语料库建设中的应用[①②]

林君峰[②]

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

Lin Junfeng

[**Abstract**] Today's large-scale construction of [illegible]

1 引 言

[illegible]

作用也受到更多关注。除了可以考察学习者口语中词汇、语法、语义、语用等方面的实际表现之外，还可以了解学习者实际的汉语语音面貌，可以对其进行声、韵、调等方面的考察与分析。"(张宝林、崔希亮，2015)南京大学、暨南大学等高校都建有汉语中介语口语语料库，预计

① 本文受教育部哲学社会科学研究重大课题攻关项目"全球汉语中介语语料库建设和研究"资助(批号12JZD018)。

② 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

享时,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。

营费用,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

数据安全也更有保障。

在国内互联网上已有云 Web 服务器、云存储、云语音等多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

误标注、建立数据库、编写检索程序等几个环节,其中转写和标注的人工耗时最长。书面语语料的标注目前还只能"在总体上采用'人标机助'的标注方式"(张宝林,2013)。口语语料的标注也类似。但现在汉语语音识别技术已有了很大的发展与进步,口语音频材料的转写如

① http://yuyin.baidu.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[①]。

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）。

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，考生跟读 1 个句子；

（9）引导语朗读，考生跟读第 5 个句子（与第 5 份小文件内容相同）；

① 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) 部分运到,但不成句,需重播录音订正较多内容的。

(4) 基本不成句,大部分内容必须重听录音由人工转录的。

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

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

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)	BD	BC
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 1 的准确度等级统计①来看，HSKK 高级部分的大部分句子都在 B 级及以上(合计 [illegible]

例 9：可是两个人给一切物件打电话的时候。（可是两个人给汽油站打电话的时候。）

① 全部样本识别结果和人工听录文本下载：http://www.hanyu123.cn/html/yuliaoku/。

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偏误标注有所帮助。

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’”，“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，便于以后对语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。

运行环境：一台云 Web 服务器，上传时接入即云存储和云语音开放平台。云存储使用七牛云存储[①]，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

① 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

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).
[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.