doi:10.3969/j.issn.1673-4785.2009.03.013

利用人类计算技术的语音语料库标注方法及其实现

沈映泉[1], 刘勇进[1], 蔡 骏[1,2], 史晓东[1]

(1.厦门大学 智能科学与技术系, 福建 厦门 361005; 2.Groupe Parole, LORIA-CNRS & INRIA, BP 239, 54600 Vandoeuvre-les-Nancy, France)

摘 要: 提出一种基于人类计算的语音语料库标注方法. 该标注方法的主要思路是通过一个基于Web的语言学习系统来收集由大量学习者(用户)输入的词汇标注和音标标注, 并从中选择出现概率最大的用户输入作为语料的正确标注. 为了保证通过这种人类计算方法获得的标注文本的质量, 使用了一些计算机辅助机制来校验收集到的标注的可靠性. 采用这种方法实现语音语料库标注的主要优点在于将语料库标注和语言学习相结合, 无需专门投入大量的人力来进行枯燥乏味的语料库标注工作, 从而节省了语料库标注的成本. 对这种基于人类计算的语音语料库标注技术进行了探讨, 说明了用于收集用户输入的语言学习系统的设计以及标注生成系统的设计. 系统的应用表明, 该标注方法能够有效、低成本地生成语音语料库的词汇标注和音标标注.

关键词: 语音语料库标注; 人类计算; 分布式知识获取; 基于Web的语言学习

中图分类号: TP39 文献标识码: A 文章编号: 1673-4785(2009)03-0270-08

Method and implementation of transcribing speech corpora based on human computation

SHEN Ying-quan[1], LIU Yong-jin[1], CAI Jun[1,2], SHI Xiao-dong[1]

(1.Department of Cognitive Science, Xiamen University, Xiamen 361005, China; 2.Groupe Parole, LORIA-CNRS & INRIA, BP 239, 54600 Vandoeuvre-les-Nancy, France)

Abstract: A new method is proposed for generating transcriptions of speech corpora based on human computation. The method depends on collection of orthographic transcriptions and phonetic transcriptions from a large number of users by using a Web-based language learning system and choosing commonly-used labels as the transcriptions of the speech corpora. In order to guarantee the quality of transcriptions, some computer-aided mechanisms are also used to verify the collected transcriptions. This method combines speech data transcribing with language learning and cuts down the cost of transcribing corpora effectively. The technology of human computation-based speech corpora transcribing and the detailed design of language learning system have been discussed. transcriptions generation system has also been expatiated in this article. The application of system shows that this method is an effective and economical way to generate orthographic and phonetic transcriptions.

Keywords: speech corpora transcription; human computation; distributed knowledge acquisition; Web-based language learning

在语音识别系统的开发中, 对语音语料库进行正确的词汇标注(orthographic transcription)和音标标注(phonetic transcription)是建立有效的语音模型和语言模型的必要条件. 然而, 为大规模语音语料库添加词汇标注和音标标注是一项需要投入大量人力、物力资源的任务. 由于现有的语音识别系统无法实现语音语料库的自动标注, 故添加词汇和音标标注往往只能通过手工标注来完成. 不论是进行词汇标注还是音标标注, 其本质都是将与语音集合对应的标注信息添加到语料库中. 这样的语音标注任务在信息添加的内容

收稿日期: 2008-07-02
基金项目: 国家留学基金资助项目(2006104705); 福建省自然科学基金资助项目(2006J0043); 厦门大学"985工程"二期信息创新平台资助项目(0000-X07204).
通信作者: 蔡 骏. E-mail: Jun.Cai@ulb.ac.be; Jun.Cai@loria.fr

和形式上与标注图像信息 (labeling images)是类似的. 因此, 图像标注的一些实现方法, 比如采用基于人类计算 (human computation)的网络游戏[1]来产生图像标注的技术, 完全可以被借鉴来解决语音语料库标注的问题. 本文据此提出了一个采用人类计算技术的 Web语言学习系统, 该系统将语音语料库的标注任务和英语学习的教学过程结合在了一起, 从而在分布式知识获取的基础上实现语音语料库的标注. 尽管这个 Web语言学习系统与图像标注的网络游戏一样, 都采用了人类计算技术, 但二者在为用户提供的服务方面存在很大的不同. 后者的服务给用户带来的只是游戏的乐趣, 而 Web语言学习系统则为英语学习者提供了一个练习英语听力理解和训练英语发音的学习平台, 用户通过这个平台获得的是一个语言学习和训练的环境.

1　语音语料库的标注

在语音识别领域, 通常需要对语音语料库进行词汇标注和音标标注, 这两类标注是训练声学模型和语言模型所不可或缺的[2-4]. 此外, 这 2种标注在其他领域也有着重要的应用, 例如为听力残障人士提供视频字幕, 以及对音频或音 /视频节目进行基于内容的搜索等等. 由于语音识别系统的质量在很大程度上取决于在识别引擎建模过程中是否有足够多的精确标注的语音语料, 因此, 对大规模语音语料进行高质量的词汇标注和音标标注在语音识别系统的开发中是一个十分重要的环节. 由于词汇标注只需提供给用户一个输入框, 接收用户的输入, 而后台的实现上完全和音标标注相同, 因此, 作者只论述音标标注.

为语音语料库添加标注信息的方法有手工标注和自动标注 2种. 下面分别介绍这 2种方法的特点及其在应用中面临的困难.

1.1　手工标注

手工标注由受过专门的语言学训练的专家来完成, 因此它直接从人类专家那里获取语言学知识. 虽然手工标注可借助如 Transcribe和 WinSnoor等一些软件工具来完成, 但对于标注者来说, 标注大型的语音语料库是一项枯燥乏味、费时费力的机械性劳动; 因此在标注过程中容易出错. 为了保证标注的质量, 通常需要由一组标注者对所有的标注文本进行交叉校验和核查, 以纠正标注中存在的错误. 这意味着在大型语音语料库标注项目的实施过程中要投入相当大的人力资源, 整个工程往往耗资巨大, 手工进行音标标注尤其如此. 正是由于广泛存在的资金投入不足, 手工标注方法一般只能用来标注小型语料库 (如 TIMIT)或者大型语料库中的一小部分. 这就导致了在开发各种语言的高性能语音识别系统时, 常常面临着缺乏高质量标注的大型语音语料库的难题.

1.2　自动标注

为了克服手工标注大型语音语料库存在的难题, 人们开发出了许多可对语音语料库自动添加词汇标注和音标标注的方法. 对各种自动标注系统, 有一个基本的质量要求, 那就是自动生成的标注应具有足够高的准确度, 使其能用于声学模型和语言模型的训练.

通常可以用自动语音识别系统 (automatic speech recognition, ASR)来生成词汇标注. 此外, ASR系统也可以用来自动生成音标标注, 例如可采用神经网络、单音子或三音子声学模型来标记和分割自然口语语音 (spontaneous speech)的音素序列[5-6]. 虽然对于新闻广播的标准朗读语音来说, 目前的 ASR系统已经能够达到超过 90%的词汇识别准确率[7-8]和 80%左右的音素识别准确率, 但这样的应用还远未达到令人满意的程度. 自动生成的词汇标注和音标标注中散布着比例相当大的错误成分, 还需要由人类标注者逐个词、逐个音标地仔细检查和校对, 以保证标注的可用性. 因此, 目前的自动语音识别系统应用并没有真正解决手工标注枯燥乏味、费时费力的问题. 另一个更严重的问题是目前技术水平的 ASR系统要求用户发音清晰、语速稳定, 而且发音和语法都必须是正确的. 然而在现场新闻报道当中, 往往存在大量的不正规的语音, 比如随意的发音、不完整的词汇、语音中的停顿、迟疑, 以及不时出现的语速变化等等. 这些语音现象的存在使得自然口语语音的自动标注变得十分困难. 自然口语语音的识别率, 特别是词汇的识别率是比较低的 (一般低于 80%)[9-11]. 因此, 由 ASR系统生成的自然口语语音的词汇和音标标注集合无法被用作可靠的语料库来建立语言模型和声学模型.

有一些音标自动标注系统通过查找发音词典的方法将词汇标注文本映射为它们的发音音标[3]. 发音词典由不同的词汇及其对应的发音组成. 这种方法的应用有一个先决条件, 那就是语音的词汇标注已经存在. 对于有多种不同发音, 或者有多种口音变化的词汇, 这种查找发音词典的方法往往难以奏效. 对于多音词, 尽管可以建立一个准确的发音词典, 词典中同时列出它的所有发音; 但是在将一个多音词映射到其发音的过程中, 发音词典本身无法提供如何根据上下文来为多音字选择对应发音的规则. 另外, 许多人名

和专有名词往往没有包含在发音词典中,因而无法通过查找词典为它们生成音标标注.这些原因导致了词典查找方法生成的音标标注集的准确率偏低,从而无法满足训练高性能声学模型的需要.

2 人类计算及其应用

在计算机科学中,将计算过程中的某些步骤或算子交给人类计算者,由人类来完成这些计算功能,这样的技术称为人类计算[12].

在传统的计算中,人向计算机提交一个问题的形式化描述,然后从计算机得到一个或多个解,由此来解释问题.但是在人类计算中,人和计算机的角色恰好相反:计算机要求一个或一群人来解决一个问题,然后将他们的解收集起来再进行解释和整合.人类计算的一般方法是将计算过程中的某些算子的运算交给用户解决,利用人类计算者本身所具有的处理能力来解决那些目前还没有可用的自动计算算子的问题,如自然语言的处理.这也体现出了人类计算的基本思想:如果能够很好地利用人类自身所具有的处理能力,那么很多计算机无法解决的问题就可以通过人类的计算、处理而得到满意的解决[1,13].

目前,人类计算技术已经成功地运用到交互系统的设计上,例如通过网页收集广大用户的最具一般性的知识[1,13-17].ESP游戏就是其中的一个典型例子.为了从网络上数以百万计的图片中找到合适的图片,有些应用软件,如图片搜索引擎和视觉残障人士的图片阅读辅助软件,要求图片附有准确的文字描述.但是目前除了手工标注外还没有一种别的方法可以用来为图片生成准确的文字描述,而当前的计算机视觉技术中也没有一种别的通用有效的方法可以用来确定图片的内容.然而 ESP游戏却很好地解决了这个问题,它将用户希望享受游戏乐趣的心理和图片标记任务有机地结合起来,利用人类计算技术为每幅图片生成有意义的文字描述.实际上,ESP游戏使用了人脑来代替电脑实现分布式计算,这类游戏一般称为“有目的的游戏”(game with a purpose,GWP).

3 系统的设计

受到 GWP思想的启发,作者设计了一个基于人类计算的 Web语言学习系统来解决在语料库标注过程中遇到的难题.GWP常受人诟病的一个不足之处在于用户花费大量的时间在游戏中,除了体验游戏的乐趣外没有其他的收获,无偿地为 GWP的开发者提供了人类计算服务.相比之下,设计的 Web语言学习系统不同于纯粹的网络游戏,它将语料库的标注任务和英语学习过程结合了起来,为用户提供了一个在线学习英语的服务平台,使得用户能通过练习注音和听写来训练自己的英语发音和听力理解.系统将用户在练习过程中输入的音标标注和听写文字收集起来,由此产生语料库的标注.

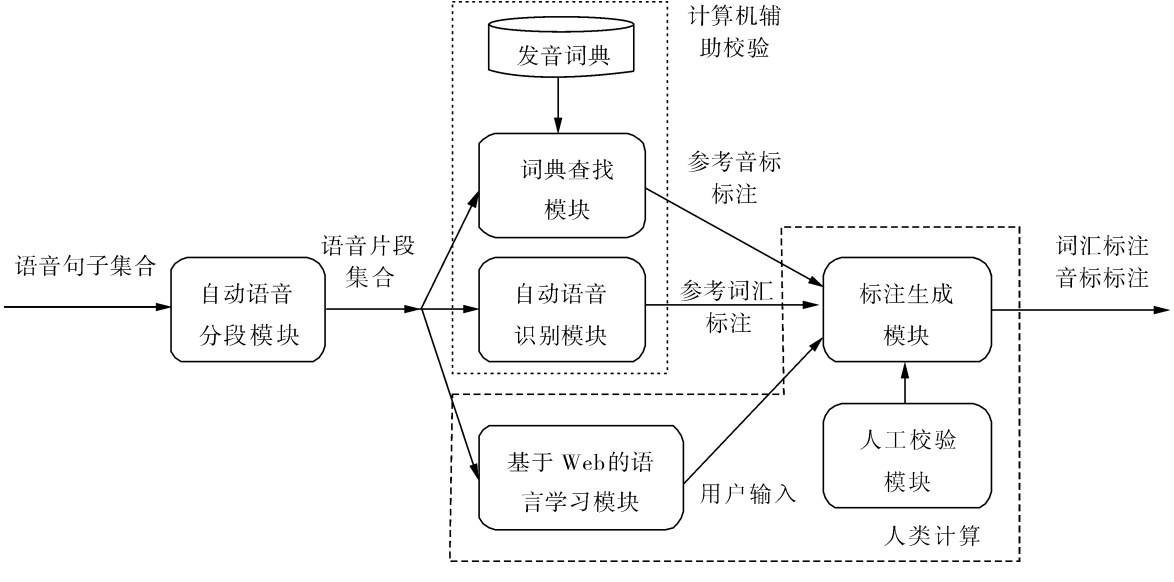

图 1 基于人类计算的 Web系统的结构

Fig. 1 Framework of the human computation-based Web learning system

图 1描述了整个系统的框架,基于 Web的语言学习模块是系统的核心,它利用人类计算技术来标注语料库.系统以语音句子集合作为输入,经过自动语音分段模块的分割后得到语音片段集合.每个语音片段被分别送到 Web学习模块、自动语音识别模块和词典查找模块.自动语音识别模块和词典查找

模块生成语音片段的参考标注文本. 在 Web学习模块, 系统将语音片段播放给用户, 当用户对语音片段进行标注并提交后, 系统收集用户输入的标注串并为用户播放下一个语音片段. 最后, 系统通过标注生成模块, 对用户输入和参考标注文本进行一定的比较处理后, 输出语料库的标注集. 在这个过程中, 系统还使用了人工校验模块对用户的输入进行校验. 下面对整个系统设计进行详细地描述.

3.1　用户界面

基于 Web的语言学习模块为用户提供了一个英语学习平台. 用户通过 Web页面上的音频播放器播放英语句子及其各个片段来练习英语发音和听力理解. 同时, 该模块收集用户输入的标注串并将它们存入 XML文件集合中 (XML文件的细节将在后面给出). 在收集到大量的用户标注串后, 系统利用人类计算机制为每个语音片段生成对应的词汇标注和音标标注.

图 2所示的是英语发音训练的网页 (http://59.77.21.117:8080/humanComputation/jsp/english.jsp). 用户的任务是收听播放的语句或语音段, 然后输入对应的音标符号串. 网页上面的音频播放器可以播放、重复播放每个语音片段. 播放器左边的文本框用来显示当前正在播放的语音段所在句子的词汇文本, 为用户的语音理解提供一定程度的提示和辅助. 如果当前播放的语音片段在系统中尚没有对应的词汇标注文本, 那么文本框中不显示任何内容.

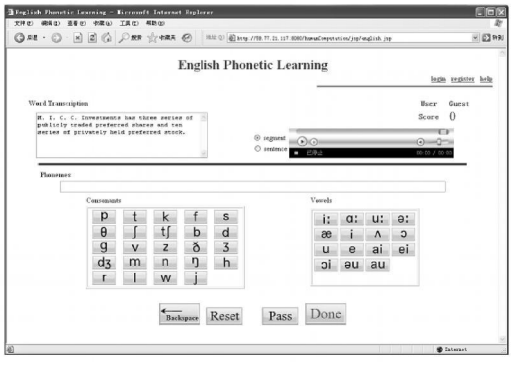

图 2　语音学习的网页界面

Fig. 2　The Web page for phonetic learning

播放器和文本框的下面是一个输入框, 用来显示用户输入的音标符号. 在系统中, 采用的是 CMU的音标, 由 15个元音音标和 24个辅音音标组成. 为了便于用户使用, 将这 39音标表示成对应的国际音标并由此构成页面上的 1个特殊的键盘, 用户只能通过鼠标点击该键盘上的按键来输入音标符号. 这样的设计将用户的输入限制在音标字母集合中, 保证用户不会输入非法的音标符号. 同时, 为了帮助用户准确、熟练地掌握每个音标的发音, 在这个特殊键盘上增加了音标发音功能. 当用户双击键盘上的一个按键时, 系统就会播放对应音标的标准发音. 键盘的下面还设置了一些按钮, 通过这些按钮用户可以选择跳过一个难懂的语音段或者向系统提交自己输入的音标串.

3.2　自动分段

系统的输入是语音句子的集合, 这些句子的长度大多超过 10 s. 对于用户来说, 标注 10 s的句子是比较困难的. 为了降低任务的难度, 使用自动语音分段模块将输入的语音句子分割成大约 2 s的语音片段.

当前系统使用的语音文件信噪比比较大, 因此使用了短时能量噪音检测技术来实现语音的自动分段[18]. 通过计算每个语音帧的短时能量值来确定该语音帧是语音还是静音, 当检测到连续 100 ms的静音时就可以确定一个语音段的端点. 通常情况下, 每个句子的开头和结尾都会有一小段的静音. 为了防止分段模块生成含有较长静音时间的语音片段, 做了一些额外的处理来保证生成的每个语音片段中至少包含一定数量的 (比如 50个) 有效语音帧. 如果句子最后一个语音片段所包含的有效语音帧太少的话, 那么模块就把该语音片段和它的前一个语音片段结合起来. 这样就保证了用户不会听到只包含 1个或 2个音节的语音片段.

3.3　计算机辅助校验 (computer-aided verification)

对于一个用户输入串, GWP(如 CMU的 ESP和 CYC的 FACTory等) 对在该输入串上达成一致的用户数量进行统计从而确定它的质量. 尽管这种机制能够有效地收集一般的事实和知识, 但是这种基于 Web的游戏本身无法保证其收集到的信息和知识是完全正确的. 在某些情况下, 尽管收集到的信息是许多人的共识, 但其实是共同的错误. 例如对于语料库标注任务来说, 如果一群学生受到的发音教育是不正确或者不准确的, 那么这些学生输入的标注很可能是一致错误的. 为了防止系统将用户的共同错误输入接纳为语料的标注, 引入了自动语音识别 (ASR)模块和词典查找模块来对用户的输入进行校验.

使用的 ASR模块是在 WSJ0的 CD1、CD2和 CD3的语料数据基础上建立起来的, 采用 WSJ0的 CD14上的 Nov'92 ARPA WSJ Test Set的 330个语句进行测试. 该模块在词汇级和音素级上识别的准确率分别为 94%和 62.51%. 播放给用户的每个语音片段同时也传给 ASR模块进行识别, 生成其对应的识别词汇串, 这里将其称为参考词汇标注. 系统用 ASR模块生成的参考词汇标注来校验用户输入的

词串. 对同一语音片段, 系统将用户输入的词汇串和参考词汇标注进行比较, 通过编辑距离[19]来计算用户输入词汇串的错误率. 定义用户输入串和参考词汇标注之间的一致性如下:

$$C = 1 - R_{error}. \quad (1)$$

式中: C表示一致性, R_{error}表示词错误率. 显然, 如果用户输入的是正确的词串, 那么该词串与参考词汇标注之间的一致性就比较高; 反之一致性就比较低. 因此, 词串的一致性可用来衡量用户输入词串的质量. 在当前的系统中, 一致性值低于 0.4的词串将被系统拒绝, 这使得低质量的词汇标注不会被系统接纳. 不论这样的标注是多少用户的共识, 这样的词汇标注也不会出现在人类计算的结果当中.

词典查找模块的作用和 ASR模块类似, 不同的是词典查找模块生成的是语音片段的参考音标标注. 词典查找模块以 ASR模块生成的参考词汇标注做为输入, 通过查找发音词典生成参考音标标注. 和词串的校验一样, 系统将用户输入的音标串和参考音标标注进行比较, 计算二者的一致性; 从而对用户输入的音标串进行评价, 以滤除用户的低质量音标串输入.

3.4 标注文本的存储

在系统实现中, 语音数据以语音句子和片段的形式存储. 为每个语音句子生成一个 XML文件, 用以存储该语音句子的标注文本和其他信息. XML文件的模式如下:

```
<? xml version="1.0" encoding="UTF-8"? >
<ANNOTATION>
  <UTTERANCE> speech file</UTTERANCE>
  <LENGTH> number of seconds</LENGTH>
  <TEXT> orthographic transcription of the sentence
  </TEXT>
  <SAMPLINGRATE> 16 <SAMPLINGRATE/>
  <WORDLENGTH> 16 </WORDLENGTH>
  <ENDIANNESS> little endian</ENDIANNESS>
  <NUMBER_SEGMENTS> n
  </NUMBER_SEGMENTS>
  <SEGMENT> segment01
    <FILENAME> segment file</FILENAME>
    <START_TIME> start point</START_TIME>
    <SEG_LENGTH> number of seconds
    </SEG_LENGTH>
    <LABEL>
      <WORD_LABEL> word level annotation
      </WORD_LABEL>
      <PHONE_LABEL> phone level annotation
      </PHONE_LABEL>
    </LABEL>
    <ANNODATA> annotation01
      <WORD_LABEL> word transcription
      </WORD_LABEL>
      <WORD_CONFIDENCE> m
      </WORD_CONFIDENCE>
      <PHONE_LABEL> phonetic transcription
      </PHONE_LABEL>
      <PHONE_CONFIDENCE> m
      </PHONE_CONFIDENCE>
    </ANNODATA>
    …
    <ANNODATA> annotation20
      <WORD_LABEL> word transcription
      </WORD_LABEL>
      <WORD_CONFIDENCE> m
      </WORD_CONFIDENCE>
      <PHONE_LABEL> phonetic transcription
      </PHONE_LABEL>
      <PHONE_CONFIDENCE> m
      </PHONE_CONFIDENCE>
    </ANNODATA>
  </SEGMENT>
  …
  <SEGMENT> segment10
    <FILENAME> segment file</FILENAME>
    <START_TIME> time of start point
    </START_TIME>
    <SEG_LENGTH> number of seconds
    </SEG_LENGTH>
    …
  </SEGMENT>
</ANNOTATION>
```

每个 SEGMENT标签标记一个切割出来的语音片段, 存储其相关信息. SEGMENT标签都包含 5个子标签, 其中 FILENAME, START_TIME和 SEG_LENGTH分别用于存储语音片段的文件名、语音片段的开始时间和持续时间, 而子标签 LABEL中的 WORD和 PHONE分别存储 ASR模块生成的词串和词典查找模块生成的音标串. 另外, 用户输入的标注及其对应的置信度值存放在子标签 ANNODATA中. 其中置信度值 (confidence value)用于反映用户输入标注的普遍性和一致性. 对于每个语音片段,

XML文件中存储其 20个不同的标注文本.

3.5　标注文本的生成

收集到大量的用户输入后, 系统利用人类计算技术为所有的语音片段生成词汇标注和音标标注. 下面仍然采用发音训练以及音标标注的生成过程为例来描述.

用户登录系统之后, 系统在语料库中随机选择出一个句子. 如果该句子的词汇标注已经存储在对应的 XML文件之中, 那么系统会在文本框显示出该词汇标注文本以帮助用户理解语音. 然后系统为用户播放该句子的一个语音片段并等待用户的反应. 如果用户输入一个音标串并且按下 "Done"按钮, 那么系统就使用下面的算法处理用户提交的音标串.

```
Algorithm: HC-based Transcribing
INPUT: A speech segment and the corresponding XML
       file; An input phoneme string
OUTPUT: A score of the input
PROCEDURE
  BEGIN
    Compare the input string with its reference string
      and compute the consistency between them;
    IF (the consistency < 40%)
      Discard the input string;
      Return a low score (say 1 point);
    ELSE
      IF (the input string is new for the segment
        and there is a free slot in the XML file)
          input string → XML file;
          Return an adequate score;
      END IF
      IF (the input string = a previpusly stored string)
          Increase the confidence measure of
          the stored string;
          Return a high score according to the confi-
                dence measure;
      END IF
      IF (the consistency > minimum confidence
          value of the segment)
          Delete the string with the minimum
          confidence;
          input string → XML file;
          Return an adequate score;
      END IF
      Discard the input string;
      Return an adequate score;
    END IF
  END
```

算法中, 与音标串相关联的置信度值起着重要的作用, 它对认同该音标串的不同用户进行记录. 如果某个音标串是第一次被输入的话, 那么它的置信度值被初始化为由式 (1)计算得到的一致性值. 此后, 该音标串每被输入 1次, 它的置信度值就加 1. 因此, 输入同一个音标串的用户数目越多, 那么这个音标串的置信度值就越大. 一旦语料库中所有的语音片段都被大量的用户反复地标注后, 那么就可以通过基于人类计算的思想来为语音片段确定最好的音标标注: 对于每个语音片段, 选择具有最大置信度值的音标串作为该语音片段最好的标注.

然而, 置信度值只是用来反映标注串的普遍性和一致性, 它无法绝对地保证标注串的质量. 如果某个标注串存在共同错误, 而它的一致性值恰好高于阈值 (40%), 那么它将被存入 XML文件中. 当这个标注串被大量的用户输入后, 它的置信度值就会比其他标注串高, 最后这个带有共同错误的标注串将被选为语音片段最好的标注.

基于人类计算的技术其本身无法摆脱在最后生成的标注集合中存在带有共同错误的标注的情况. 为了解决这个问题, 在系统中引入了人工审核模块以帮助受过训练的专家对生成的标注进行有选择地检查, 将存在问题的标注从 XML文件中删除掉. 另一方面, 如果某个用户在标注的过程中经常犯错误, 那么可以通过人工审核模块删除掉该用户输入所有标注. 将人工审核和人类计算结合在一起, 可以有效地保证最终生成的标注集具有较高的质量.

4　实验统计结果

目前作者的 Web系统使用 Nov'92 ARPA WSJ Test Set中的 330个语音句子作为测试语料, 使用自动语音分段模块对这些数据进行切割后, 一共生成了 920个语音片段. 测试过程中, 共有 19个学习者使用系统对这些语音片段进行标注, 最后收集到 1 900个标注结果.

在对这 1 900个标注结果进行处理之前, 分别定义语音片段 i的召回率 R_{seg}和准确率 P_{seg}.

对于某个由自动分段模块切割所得的语音片段 i, r表示语音片段 i经过 ASR模块识别后所得参考音素串, s表示用户对语音片段 i进行标注后所得音素串. $L(str)$表示串 str中音素个数. $D_L(s,t)$表示 s和 t的编辑距离 (Levenshtein distance), 它表征了字符串 s和 t的差异度. s和 r中相同的音素个数为

$$S(i,t_s) = L(t) - D(i,s_t).$$

由此,定义语音片段 i 的召回率 R_{seg} 和准确率 P_{seg} 如下:

$$R_{seg}(i) = (S(i,t_s) / L(s)) \cdot 100\%,$$

$$P_{seg}(i) = (S(i,t_s) / L(t)) \cdot 100\%.$$

下面是对收集到的 1 900个标注结果的处理方法:对于音素串 t,如果其对应的语音片段 i 的召回率或者准确率低于 40%,则丢弃音素串 t;如果同一个语音片段有 2个或 2个以上的标注结果,那么保留准确率较高的标注结果.

经过处理后,最后得到 588个带有用户标注的语音片段,计算总的召回率和准确率如下:

$$R_{seg} = (\sum S(i,t_s) / \sum L(t)) \cdot 100\%,$$

$$P_{seg} = (\sum S(i,t_s) / \sum L(s)) \cdot 100\%.$$

最终计算出来总召回率和准确率为

$$R_{seg} = 72.063\%,$$

$$P_{seg} = 72.186\%.$$

这说明经计算机辅助校验模块筛选后的用户输入与参考标注串之间有较高的一致性,保留下来的用户输入有较好的质量,在此基础上生成的音标标注的准确性是可信的.

5 结束语

文中提出了一种基于人类计算的方法来完成语音语料库的标注工作.通过设计一个 Web语言学习系统来收集学习者输入的词汇标注和音标标注,然后利用人类计算技术生成语料库的标注集,同时由受过训练的专家通过人工校验模块对生成的标注集进行人工审核以进一步保证其质量.该 Web系统的一个特点是将语料库标注任务和英语的学习过程结合在一起.当前的 Web系统使用 WSJ0语料库中的一部分作为测试语料数据.系统在测试用户较少的情况下在较短的时间内完成了对测试语料数据的标注,最后的统计显示了系统生成的标注集有较高的准确率.这也初步证明了基于人类的方法能够有效、低成本地完成语料库标注工作.作者相信,在大量用户使用该系统的基础上,系统可以收集到大量的用户标注文本从而为语料库生成高质量的标注集.

当然,目前文中的 Web系统还存在很多不足,需要进一步改善以便提高系统的性能和功效.首先,应该推广该 Web语言学习系统的应用,以便对系统的功效做出检验和评价.其次,人类计算是以大量的用户参与为基础的,因此应该在系统中加入更多的语言学习功能以便更好地帮助并吸引更多的学习者.比如在语言学习模块中实现更多的交互,让用户能够更容易地掌握英语的发音.此外,系统的鼓励机制同样的是一个急需完善的部分.例如,与其他的一些付费英语学习网站合作,允许用户通过学习之中获得的游戏积分,兑换一些付费资料,从而达到激励用户的目的.最后,目前的系统并不是适合口语语音的标注,因此,在后继的研究之中必须添加非语言现象的标注规则.

同时还需要进行一些相关的后继研究以扩展基于人类计算的方法在其他方面的应用.比如当前系统中的自动语音分段方法还不能够完全适用于英语或法语的语音.可将长的英语或法语句子的分段工作交给学习者来完成,即采用人类计算的方法来最终确定这些语音句子的最好分割.还可以利用这个框架来转换不同的语料库,比如通过用户输入转换文本或者录制转换语音将英语语料库转换为汉语语料库.通过这种方法,可以收集到更多不同说话者的自然语音记录,从而建立更大的语料库.

参考文献:

[1] AHN L von, DABBISH L. Labeling images with a computer game[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Vienna, Austria, 2004: 319-326.

[2] BIRD S, LIBERMAN M. A formal framework for linguistic annotation[J]. Speech Communication, 2001, 33(1/2): 23-60.

[3] YOUNG S J, EVERMANN G, GALES M, et al. The HTK book (for HTK Version 3.4)[EB/OL]. [2008-06-20]. http://htk.eng.cam.ac.uk/prot-docs/htk_book_shtml.

[4] DEMUYNCK K, LAUREYS T, GILLIS S. Automatic generation of phonetic transcriptions for large speech corpora[C]//Proceedings of the 7th International Conference on Spoken Language Processing. Denver, USA, 2002: 333-336.

[5] SCHIEL F. Automatic phonetic transcription of non-prompted speech[C]//Proceedings of 1999 International Conference of Phonetic Sciences. San Francisco, USA, 1999: 607-610.

[6] CHANG S, SHASTRI L, GREENBERG S. Automatic phonetic transcription of spontaneous speech (American English)[C]//Proceedings of the 6th International Conference on Spoken Language Processing. Beijing, 2000, 4: 330-333.

[7] CHEN S S, EIDE E, GALES M J F, et al. Automatic transcription of broadcast news[J]. Speech Communication, 2002, 37(1/2): 69-87.

[8] CHAN H Y, WOODLAND P. Improving broadcast news transcription by lightly supervised discriminative training[C]//Proceedings of 2004 IEEE International Conference

on Acoustics, Speech, and Signal Processing. Montreal, Canada, 2004, 1: 737-740.

[9] KATO K, NANJO H, KAWAHARA T. Automatic transcription of lecture speech using topic-independent language modeling[C]//Proceedings of the Sixth International Conference on Spoken Language Processing. Beijing, China, 2000, 162-165.

[10] BACCHIANI M. Automatic transcription of voicemail at AT&T[C]//Proceedings of 2001 IEEE International Conference on Acoustics, Speech and Signal Processing. Salt Lake City, USA, 2001, 1: 25-28.

[11] HAIN T, BURGET L, DINES J, et al. The 2005 AMI system for the transcription of speech in meetings[J]. Lecture Notes in Computer Science, 2006, 3869: 450-462.

[12] KOSORUKOFF A. Human-based genetic algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2001, 5: 3464-3469.

[13] AHN L von. Human computation[EB/OL]. (2006-07-26)[2008-06-20]. http://video.google.com/videoplay?docid=-8246463980976635143.

[14] SINGH P, LIN T, MUELLER E T, et al. Open mind common sense: knowledge acquisition from the general public[J]. Lecture Notes in Computer Science, 2002, 2519: 1223-1237.

[15] GENTRY C, RAMZAN Z, STUBBLEBINE S. Secure distributed human computation[C]//Proceedings of the 6th ACM Conference on Electronic Commerce. New York, USA: ACM, 2005: 155-164.

[16] AHN L von, KEDIA M, BLUM M. Verbosity: a game for collecting common-sense facts[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, USA: ACM, 2006: 75-78.

[17] AMAZON.COM, Inc. Amazon mechanical turk[EB/OL]. [2008-06-20]. http://www.mturk.com.

[18] DONG Enqing, LIU Guizhong, ZHOU Yatong, et al. Voice activity detection based on short-time energy and noise spectrum adaptation[C]//Proceedings of the 6th International Conference on Signal Processing. Beijing, China, 2002, 1: 464-467.

[19] BUNKE H. On a relation between graph edit distance and maximum common subgraph[J]. Pattern Recognition Letters, 1997, 18(9): 689-694.

作者简介：

沈映泉，男，1984年生，硕士研究生，主要研究方向为语音情感识别、自然语言处理.

刘勇进，男，1984年生，硕士研究生，主要研究方向为自然语言处理.

蔡　骏，男，1966年出生，副教授，博士. 布鲁塞尔自由大学(ULB)图像、信号和远程通信实验室研究员. IEEE Computer Society, IEEE Signal Processing Society会员，International Speech Communication Association会员. 主要研究方向为自动话语识别、计算机语音处理，在自动话语识别的实时计算和人类语音的 Articulatory Modeling等方面进行了深入的研究. 参加与主持科研项目 20项，发表学术论文 30余篇.