中图分类号:TN912.3 文献标识码:A 文章编号:1009-2552(2006)08-0017-04

基于 Web 网页语料构建动态语言模型

李雪涛,文茂平,杨 鉴

(云南大学信息学院,昆明 650091)

摘 要:为语音识别系统构建语言模型,首先要进行语料准备,语料来源决定语言模型的性能。Web 网页中涵盖了各种最新的语言现象,为语料准备提供了最多样化的资源。但 Web 网页中语义完整字串通常夹杂在格式、标记、广告等无用字串中。首先介绍语言模型的训练算法和更新方法,继而提出一种从 HTML 文档提取用于训练语言模型的语义完整汉字字串的算法,最后给出语料提取实验结果、语言模型训练结果和语言模型的动态更新结果。为基于 Web 网页语料动态更新语言模型提供了一个完整的解决方案。

关键词:语言模型;语料库;信息提取;动态更新

Updating language model based on training text from Webs

LI Xue-tao, WEN Mao-ping, YANG Jian

(**School of Information Science and Technology, Yunnan University, Kunming** 650091, **China**)

Abstract: A statistical n-gram language model (LM) is used to predict each language symbol in the sequence given its n-1 predecessors. The first stage of constructing an n-gram LM for speech recognition system is gathering training text set. Web is a vast repository of information and a very important resource of the training text set for updating LM. However, the HTML documents downloaded from the Web include a lot of redundant text for training LM, such as format, tags and advertisements. In this paper, a new algorithm to automatically extract the Chinese training text from HTML documents is introduced. Based on the algorithm bout 93MB training text set is extracted from Webs and a baseline 3-gram LM is constructed using this text set. To verify that the updating LM based on the Web is effective, another training text set, about 14MB and from Webs, is used to update the baseline LM. In addition, the perplexity and the OOV (Out of Vocabulary) of the baseline LM and the updated LM are estimated, respectively.

Key words: language model; Chinese training text; extracting information; updating

0 引言

语音识别,是新一代人机系统的核心技术之一。其目的是要准确地理解语音所蕴涵的语义,其基本任务是根据观测到的声学信号推测出最有可能的词序列,这一过程可用下面的公式描述:

$$W^{*} = \arg\max_{W} P(W \mid O) = \arg\max_{W} P(O \mid W)\, P(W) \quad (1)$$

其中,O 是从声学信号中提取的特征向量,W 是待确定的汉字串或词串。$P(O \mid W)$ 为在语言序列 W 产生特征向量 O 的条件概率,它是声学模型的建模基础,$P(W)$ 是对应的语言序列产生概率,是 W 独立于语音特征矢量的先验概率,由语言模型决定[1]。在连续语音识别系统中语言模型必不可少,特别是一些同音字必须通过上下文才能确定,语言模型直接影响识别系统的性能。

为了构建能够覆盖尽可能多的语言现象的语言模型,需要准备大量的涵盖特定领域方方面面的训

收稿日期:2006-02-09

基金项目:国家自然科学基金项目(60265001)

作者简介:李雪涛(1980-),女,2003 年毕业于东北大学,云南大学通信与信息系统专业在读硕士研究生,主要研究方向为语音识别算法。

练语料。随着 Internet 及其技术的迅速发展,Web 已经成为当今最庞大的信息库,并且这个自由的网络空间包含最新最多样化的语言现象,为语料的准备提供了最丰富的资源。但是,除了少数文本网页外,绝大多数网页的内容都是存在于半结构化的 HTML 文档中[2],为语料的提取带来很大困难。而且,用于构建汉语语言模型的训练语料应为前后词语连贯,意思完整的汉语句子(为了方便讨论,本文称这样字符串为语义完整的字串),所以 Web 页面含有很多多余信息。另外,同一网站的不同网页通常含有大量的相同的广告信息,这并不反映通常情况下人们说话的习惯和汉语使用的频率,若将这些信息也作为训练语料,会大大影响语言模型的性能。本文提出的网页语料获取方法,利用 HTML 文档的通用格式,能够迅速提取 HTML 文档中符合语料标准的部分,将其存于无格式文档中,大大提高了语料的质量和语料准备的效率。

另外,在大词汇量语音识别系统中,即使采用数以千兆计的语料来训练统计语言模型,一样存在着数据稀疏问题,而且随着时间的推移,会不断涌现出很多新词汇,不间断地用最新的网页语料对原有语言模型进行添补和更新可有效缓解数据稀疏问题,并使其覆盖最新的语言现象。如何快捷而有效地运用网络资源来构建语言模型并动态更新语言模型使它覆盖更多更新语言现象是本文的研究目标。

本文首先讨论语言模型与语料库,然后介绍从 Web 网页中自动获取语料的算法,最后给出语料提取实验结果、语言模型实验结果和语言模型的更新结果,以验证基于 Web 网页语料动态更新语言模型的有效性。

1 语言模型及其动态更新

语言模型可分为基于规则和基于统计两种。基于规则的语言模型通过专家知识总结出语法语义规则,然后利用这些规则排除声学语音层的搜索识别中不和语法或语义规则的结果。基于统计的语言模型,通过对大量文本信息的统计,提取出不同词条(子词)的出现概率及其相互关联的条件概率,其特点是数据准备一致性好,鲁棒性强,适合处理大规模真实语料[3]。本文采用统计的方法构建语言模型。

1.1 语言模型及其训练算法

建立语言模型,也就是要求出词序列的概率估计 $P(W)$,可由下式给出[4-5]:

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \cdots, w_{i-1}) \tag{2}$$

若为词的 n - gram 模型,则(2)式可被限制成:

$$P(w_1, w_2, \cdots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, \cdots, w_{i-1}) \quad n \geq 2 \tag{3}$$

基于实用上的考虑,n 值一般取 2 至 4。通常采用最大似然估计方法计算 n - gram 模型参数,即通过在给定的训练文本中统计事件发生的次数:

$$P(w_i \mid w_{i-n+1}, \cdots, w_{i-1}) = \frac{C(w_{i-n+1}, \cdots, w_i)}{C(w_{i-n+1}, \cdots, w_{i-1})} \tag{4}$$

式中 C(·)为在训练文本中给定词序列发生的次数。

HTK (Hidden Markov Toolkit) 是一个基于隐马尔可夫模型的开放源代码的语音识别工具包,由英国剑桥大学工程系机器智能实验室开发。它包括了以 C 语言为源代码的一系列库模块与工具。HTK 主要是用来构筑基于 HMM 的语音识别系统,当然也可以用于构筑其他的时序信号的 HMM。它不仅可以用于构建孤立词识别系统,还可以构建大词汇量连续语音识别系统。本文主要采用其中的 HLM (语言模型) 工具来训练和更新基于词的三元文法 (trigram) 语言模型。

1.2 动态更新语言模型

建好基本语言模型之后,为了能获得与当前的最流行的语言现象相匹配的语言模型,需要动态地更新语言模型。更新语言模型最简单的办法是把新的语料与以前的语料合并,并重新训练新的语言模型,但是这种方法没有用已训练好的基本模型,很多工作要重新做,运算量太大。本文采用另一种方法,这种方法分两个阶段:第一阶段,用新的文本数据训练一个新的模型;第二阶段,基本语言模型和新的模型以一定的权值合并成一个模型,完成更新。这种方法需要确定模型合并的权值。本文假设所有的训练文本对语言模型的贡献度相同,采用与训练文本的规模成比例的方法来确定权值,即训练文本规模大的模型则权值大。

在训练语言模型之前,先要进行语料准备工作,即构建一个能体现通常情况下词间联系和语言使用频繁状况的语料库。

2 基于 Web 的语料提取及其算法

从 Web 中获取语义完整的语料需要对 Web 进行信息提取。在 Web 信息提取领域,已经有大量的研究工作,基本可分为结构分析法、tag 分析法和机器学习法等[2]。语料提取的目的是提取具有完整语义的中文字串,排除大量重复出现的广告和链接等多余信息。目前能较好解决这一问题的方法是面向内容的提取方法,它们的目标不是提取完整的数据而是提取主题内容和感兴趣的区域。其中 Finn 等人将 HTML 文档看作是字符和标签组成的序列,在字符集中的区域提取用于网页信息分类的主题文字[6]。本文沿用将 HTML 文档看作是字符和标签组成的序列的方法,在此基础上以自动获取中文语料为目的提出 Web 信息提取算法。

HTML 文档结构可表示为字符串(text)和标签(tag)的连接,如图 1 所示。

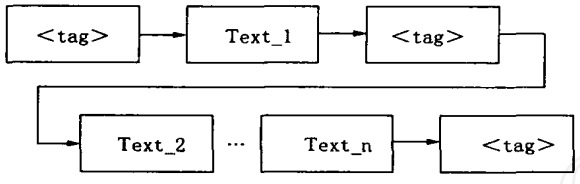

图 1 HTML 文档结构图

本文的提取算法不关心控制标签的具体含义,而是用匹配的方法找出 HTML 文本中成对出现的'<>',将'>'和下一个出现的'<'之间的内容存入缓冲区 TEXT,然后判断 TEXT 中是否含有语义完整的中文字串,若有,则将其存入缓冲区 TRAIN,若无,则不保留。然后清空 TEXT,接着用相同的策略往下搜索和处理,直到文本结束。最后,缓冲区 TRAIN 中即为所提取的语料。

通过对 HTML 网页的抽样分析得出,具有完整语义的中文字串连续汉字数较多,相比之下无关信息的汉字数较少,一般不多于 50 个汉字,例如广告信息和文章标题等。另外,TEXT 中会出现汉字和非中文字符混杂出现的情况,此时虽然中文字符总数够多,但它们是截断的,语义不完整的,因此 TEXT 中的中文字符必须占绝大多数。

基于以上的分析,得出判断 TEXT 中文本是否语义完整的两个主要条件:

条件 1:

$$\frac{C\_COUNT}{T\_COUNT} > Pc \tag{2}$$

其中,T_COUNT 和 C_COUNT 分别为 TEXT 的总字节数和中文字符字节数,Pc 为阈值,0 < Pc < 1.0。

条件 2:

$$C\_COUNT > Lc \tag{3}$$

Lc 为另一阈值,用于限制语义完整句子的中文字符字节数。

```
Initialize the result set TRAIN to be Null
HD ← The HTML document
if HD is not empty
 then repeat
      search the tag in HD
      if HD[i] is tag
       then search the following tag HD[j]
        TEXT ← the string between HD[i+1] and HD[j-1]
        C_COUNT ← the bit count of the chinese character in TEXT
        T_COUNT ← the bit count of TEXT
        C_TEXT ← the chinese character in TEXT
        if C_COUNT / T_COUNT > threshold Pc and
                               C_COUNT > threshold Lc
         then add the C_TEXT to TRAIN
        delect TEXT
      until the end of HD
```

图 2 语料提取算法

图 2 为基于以上算法思想设计的从 Web 页面中提取语义完整语料算法。

3 实验结果

3.1 语料提取实验结果

表 1 为图 2 所示算法的部分实验结果。其中,平均压缩比是语义完整无格式文本的大小与 HTML 文件大小之比。通过对部分结果进行人工分析得出,网页文本中被删除的可用文本(即语义完整文本)不超过总共可用文本的 5 %。

表 1 提取实验结果

来源 Web 网页	网页数	平均压缩比(%)
人民网	1039	2.7
新华网	649	4.2
CCTV	1344	6.0
中国新闻网	944	2.5

经大量实验分析发现,两个完整度条件的阈值的选取对结果有重要影响,可根据实验结果调整阈值的大小。适当增加阈值可保证有效删除广告、链接等多余字符,提高语料质量,但阈值过大会带来删除可用文本、减小原始语料的利用率问题。表 1 所示实验中两阈值的取值分别为 Pc = 80 %,Lc = 100,实验结果表明该组参数较好地均衡了语料质量和利

用率。

图 3 为 Web 网页实例,图 4 为用本文算法提取的与图 3 对应的语义完整文本。

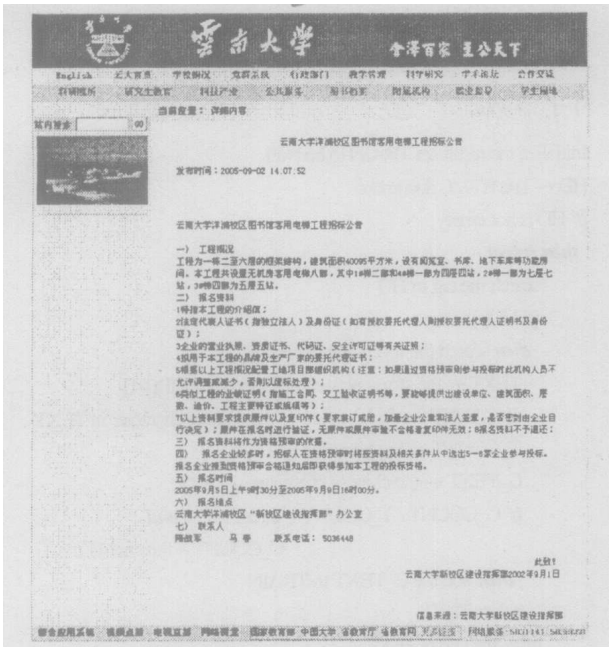

图 3 Web 网页实例

工程为一栋二至六层的框架结构，建筑面积平方米，设有阅览室、书库、地下车库等功能房间。本工程共设置无机房客用电梯八部，其中梯二部和梯一部为四层四站，梯一部为七层七站，梯四部为五层五站。根据以上工程概况配置工地项目部组织机构（注意：如果通过资格预审则参与投标时此机构人员不允许调整或减少，否则以废标处理）；类似工程的业绩证明（指施工合同、交工验收证明书等，要能够提供出建设单位、建筑面积、层数、造价、工程主要特征或规模等）；以上资料要求提供原件以及复印件（要求装订成册，加盖企业公章和法人签章，是否密封由企业自行决定）；原件在报名时进行验证，无原件或原件审验不合格者复印件无效；报名资料不予退还；四）报名企业较多时，招标人在资格预审时将按资料及相关条件从中选出一家企业参与投标。报名企业接到资格预审合格通知后即获得参加本工程的投标资格。

图 4 图 3 网页的提取结果

3.2 语言模型实验结果

采用本文算法从 Web 搜集语义完整语料,包括新闻、科技、经济、财经、体育等各个领域的文本,约 93MB,词典收词 70 262 条。然后,对其分词、注音,并转换成 HTK 工具能够处理的格式。最后,应用 HTK 语言模型训练工具构建基于词的三元文法基本语言模型,并测试它的性能。此外,再从 Web 网页搜集约 14MB 的语料,用 HTK 工具先训练生成另一个新的模型。在此基础上,基本语言模型和新的模型以一定的权值合并成一个模型,完成模型更新。实验中,用相同的测试语料测试基本模型和更新后的模型,测试语料句子数为 87 793,词数为 703 709 条,涵盖新闻、科技、经济、财经、体育、军事等六个领域的文本。测试结果如表 2 所示。其中,OOV (Out Of Vocabulary) 表示测试语料中有但模型词典中没有的词。

表 2 语言模型测试结果

	困惑度 (Perplexity)	OOV	OOV 率(%)
基本语言模型	409.7	1913	0.27
更新后的语言模型	357.8	1872	0.26

实验结果表明,经更新后的语言模型比基本语言模型性能有了适当的提高,若对原有语言模型进行多次更新,性能将更好。

4 结束语

本文从 Web 网页中提取语义完整语料,并在此基础上训练语言模型,然后搜集新的网页语料对基本模型进行更新,使语言模型能够覆盖更多的语言现象,并为自动收集规模更大、题材更广的语料,动态更新语言模型提供了一种完整的实现方案。下一步的工作是收集更多的语料,动态更新生成性能更好的语言模型,然后在语音识别系统中验证语言模型的有效性。

参 考 文 献:

[1] 蔡莲红, 黄德智, 蔡锐. 现代语音技术基础与应用[M]. 清华大学出版社,2003.

[2] Luo Xiao. Dieter Wissmann. Information Extraction from the Web: System and Techniques[J]. Applied Intelligence 2004,21:195 - 224.

[3] Jelinek F. The development of an experimental discrete dictation recognizer[J]. //Proceedings of the IEEE, 1973:1616 - 1624.

[4] Steve Yong, Dan Kershaw, et al. The HTK Book[M/OL]. (2002 - 12). http://htk.eng.cam.ac.uk/.

[5] Moore GL. Adaptive Statistical Class - based Language Modeling[D]. Ph. D thesis, Cambridge University, 2001.

[6] Aidan Finn, Nicholas Kushmerick, Barry Smyth. Fact or fiction: Content classification for digital libraries[C]. The 2nd DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin, Ireland, 2001.

责任编辑:么丽苹