863 计划中文信息处理与智能人机接口基础数据库的设计和实现①

钱跃良② 林守勋 刘 群 刘 洋 刘 宏③ 谢 萦

(中国科学院计算技术研究所 北京 100080)

摘 要 本文简单分析了中文信息处理与智能人机接口领域基础数据建设的意义以及国内外的现状,重点介绍了 863 计划中文信息处理与智能人机接口领域基础数据库的设计与实现,并对基础数据的应用和今后的工作设想进行了讨论。

关键词 中文信息处理,智能人机接口,基础数据库

0 引 言

中文信息处理与智能人机接口技术是计算机与人工智能技术的一个重要研究领域,同时也涉及到语言学、声学、光学、人体工程学等相关领域。开展中文信息处理和智能化人机接口技术的研究,对于我国信息化建设和信息产业的发展具有重要的战略意义。中文信息处理与智能人机接口技术的研究范围十分宽广,目前主要的研究内容有:自然语言处理技术(包括机器翻译、检索和自动文摘等)、文字识别技术(包括印刷体和手写体文字识别等)、语音处理技术(包括语音识别与语音合成等)、计算机视觉与图像处理技术、生物特征信息处理技术(包括指纹识别、虹膜识别、脸像识别、笔迹鉴别和声纹识别等)、多媒体技术、虚拟现实技术和多模式人机交互技术等等。

在开展上述关键技术研究过程中,其核心算法不管采用的是统计方法、规则方法,还是目前流行的基于语料库的方法,总体上讲,都需要一定规模的语料或样本对核心算法进行训练。在这里我们把这些语料和样本数据称之为基础数据。显然,基础数据库是开展中文信息处理和智能化人机交互技术研究与开发的基础。此外,基于基础数据库的关键技术评测已成为一种客观评价的重要手段。

1 国内外基础数据库的现状

正是由于基础数据库的重要性,国内外科研机构投入了大量的人力和物力来设计和制作基础数据库。目前比较有影响的基础数据库主要是语言学数据联盟(Linguistic Data Consortium,LDC[1])提供的数据库。LDC 由 DARPA 和 NSF 资助,于 1992 年在美国宾西法尼亚大学建立,目的在于建造、收集和发行语言学数据,用于语言信息处理领域的研究和发展。现在已经有 100 多所大学、公司和政府部门加盟,有各种语言学数据 220 种,涉及英、德、法、西班牙、中、日和阿拉伯文等多种语言。LDC 已经向近 700 个单位发行了数据。在 LDC 的众多数据中,包括了相当多的语音数据,覆盖了汉语、英语、德语、西班牙语、阿拉伯语等许多主要的语言。同时,语音内容极其丰富,包括:新闻广播、口语对话、对讲机对话、手机对话以及噪音环境下的对话等,这些语音数据都是在真实环境下录制的,因此大大增加了语音识别的难度,同时也反映了基础数据建设向自然方向发展的趋势。

国际上现在有很多著名的基于评测的国际会议,这些会议往往也会为参加评测者提供大量的训练数据和测试数据。例如信息检索领域的文本检索会议(TREC)[2]、日本信息检索评测会议(NTCIR)[3]、话题跟踪与检测会议(TDT)和信息理解会议(MUC)等等,这些会议都为相关技术的研究发展提供了一个可比的评测环境和开放的技术交流论坛。以 TREC 会议为例:TREC 会议每次评测都分成很多评测任务(一个任务称为一个 Track),近年来数据规模最大的评测任务是网页检索任务(Web Track),这个任务的所有数据都是从 Internet 上.GOV 结尾的网站(即政府网站)上采集下来的,因而又称为.GOV 数据集。前两年的.GOV 数据集规模为:文档数约为

① 863 计划(2001AA114010,2004AA114010)和北京市相关课题(H030130050230)资助项目。
② 男,1960 年生,大学本科,高工,研究方向:中文信息处理与智能人机接口,嵌入式系统,数字化技术。
③ 联系人,E-mail:hliu@ict.ac.cn
(收稿日期:2004-09-13)

万方数据

124.8 万个,平均文档大小为 15.2k,数据规模为 18.1G。在今年的 TREC 评测中,提出了一个新的 TB 级评测任务(Terabytes Track),这个任务是用了一个新的.GOV 数据集,称为.GOV2 数据集,数据扩大了 20 倍,其中 文档数约为 2520.5 万个,平均文档大小为 17.7k,数据规模为 426G。

基础数据库建设是一个不断更新、不断发展的长期任务。国外基础数据库建设已经形成比较成熟的机制,国内相应的工作还处于一个发展的阶段。国家 863 计划于 1990 年建立了 500 万手写汉字样本,1996 年建立了 200 人连续语音样本库。“十五”期间,国家 863 计划计算机软硬件技术主题专家组对中文信息处理和智能人机接口领域关键技术的研究发展进行了全面的部署,并设立了‘中文平台总体技术研究和基础数据库建设’的重点课题,集中支持了中文信息处理与智能人机接口领域的基础数据库(以下简称 863 基础数据库)的建设,其目的在于为相关关键技术研究和科学评测提供基础数据,避免重复劳动,促进交流与合作。基础数据库由中国科学院计算技术研究所负责,并联合了中国科学院自动化研究所、清华大学、社会科学院语言所、北京大学等多家单位共同完成。

2 基础数据库的设计与实现

按照‘需求驱动,总体规划,分段实施’的设计原则,根据关键技术研究的需要和技术发展趋势,进行基础数据库的总体规划,并按照关键技术对技术数据库需要的紧缓程度,分阶段地开展基础数据库的建设。经过多年的努力,目前 863 基础数据库已经有较大的规模和覆盖面,整个数据库的容量超过 260G,包括了以下 6 个大类的 16 个库[4]。

2.1 文字类

文字类的数据主要用于文字识别的关键技术研究和产品开发,目前主要建立了大字符集联机手写汉字识别样本库,约 400 万字的样本,具体内容如下:

字符集	人数	每人书写样本数	合计样本数
GB/T18030(不含 GB/T2312)	100 人	20 721	2 072 100
GB/T18030 全部	70 人	27 484	1 923 800

2.2 文本类

文本类的基础数据主要用于自然语言处理方面的关键技术研究与产品开发,应用对象包括机器翻译、文本分类、信息检索、信息提取等。目前文本类的基础数据库建立了两个:

(1)双语语料库

双语语料库包括汉英双语语料库和汉日双语语料库,其中汉英双语语料库包括汉英词对齐语料 10 102句对、汉英句对齐语料 302 151 句对,汉日双语语料库包括汉日句对齐语料 20 000 句对。语料既有文本语料,也有对话语料,涉及政府白皮书、政府公文、新闻、旅游、餐饮、经贸、科技文献、政治文献、法律、文学、体育、环境、艺术等领域。

(2)多语言常用词库

多语言常用词库包括以下 2 个子库:

名称	语料	规模
汉英体育词库	体育学及体育比赛规则、运动学等相关学科词汇	55 398 条
汉英旅游词库	旅游、交通、邮电、餐饮、购物等方面词汇	5200 条

2.3 音频类

音频类的基础数据主要用于语音识别和语音合成等关键技术的研究和产品开发。目前音频类数据库有 7 个:

(1)语音识别样本库

语音识别样本库共建立了 8 个子库,其数据量达到 17G,语料以新闻类为主,发音人平均男女各半,内容包括:

库名	语料内容	容量	时间长度
南方口音语音库	1 000 000 句	8.2GB	70 小时
北方口音语音库	75 000 句	5.8GB	50 小时
汉语对话语音库	10 000 句	1.1GB	8.4 小时
英语对话语音库	10 000 句	1GB	8 小时
嵌入式环境语音库	2 331 句	71.1MB	1.3 小时
信息亭通道下自适应语音库	1 000 句	46MB	40 分钟
面向奥运的英语语音库	15 000 句	900MB	7 小时
面向奥运的日语语音库	2 000 句	400MB	3 小时

(2)方言地区普通话语音样本库

方言地区普通话语音样本库是从上海、重庆、厦门和广州四个大方言区各选择了 200 人(男女各 100 人,共 800 人)进行录音,每个发音人 1~2 小时的发音时间。语料包括 220 个短语篇作为平衡语料(每个语篇不超过 50 个汉字)、160 个口语话题和 4 个城

万方数据

市常用口语方言词汇。语音数据采用 16kHz 采样,16bit PCM 方式保存。所有 800 人的录音数据均做了语音学标注,称为"粗标注"(汉字和拼音两层标注)并在每个方言点挑选出 20 人(共 80 人)的语音数据做了"精细标注"(汉字、拼音、有时间切分点的音节层、实际发音的声韵母层等)。

(3)电话语音识别样本库

电话语音识别样本库的录音人数为 700 人,年龄分布在 14 至 68 岁,来自 28 个省市、自治区;语音库包括北京本地电话数据和来自 8 省市的长途电话数据。语料包括:10 个数字和 5 个数字串、20 个最常用的单词、5 个货币金额的十进制数字读法、6 个时间(3 个日期、3 个常用时间)、8 个覆盖音节平衡的疑问句、20 个考虑音节内和音节间的覆盖和平衡的句子。语音数据采用 8kHz 采样,16bit PCM 方式保存。数据量每人大于 3M,合计大于 2.1G。

(4)语音合成数据库

语音合成数据库主要用于语音合成的研究,该库包括 4 个子库:

库名	语料		录音数据量	标注数据量
	男声	女声		
普通话 TTS 系统数据库	4491 句	6046 句	1350M	42M
普通话 TTS 系统测试数据库	1000 句	1000 句	522M	28M
普通话语调分析用数据库	0	1000 句	286M	15M
语流韵律分析数据库	50 篇	50 篇	1.6GB	2M

(5)语流韵律分析数据库

语流韵律分析数据库采集了广播电台 4 位(男女各 2 位)播音员的语音,语料为 549 个篇章,4693 个句子,其中 596 个句子为疑问句、感叹句和祈使句等,数据量为 10G,约 17 小时。

(6)中英文混读数据库

中英文混读数据库主要用于中英文混读的语音合成,数据库内容包括:汉语文本中包含字母缩写的情景语句、单个英文单词的情景语句和英文短语的情景语句。语料规模为 3000 个句子,其中有 1400 句左右中英文混合的句子、1000 句左右英文句子和 600 句左右中文句子。数据包括男声和女声两部分,每部分均含语音数据以及对语音数据检查和部分数据的标注。语音数据采用 16kHz 采样,16bit PCM 方式保存,数据量约 900M。

(7)特定场景特定领域对话语料库

该语料库采录了 5 对人(5 男 5 女)的对话语料。对话语料分成四个方面:餐饮、旅游、交通、天气。每个方面的语料有 100 个对话,其中包含少量英语对话。数据量 1.2G。

2.4 视频类

视频类数据库主要用于研究虚拟现实、3D 游戏、体育、舞蹈的训练系统等,目前建立了三个数据库:

(1)体育运动类视频数据库

体育运动类视频数据库目前主要建立了个体动作部分,包括的运动项目有:跳水、蹦床、乒乓球、羽毛球、网球、台球、体操、游泳、田径等项目,动作片断有 231 个。数据格式采用 MPEG-1,其中 3 个跳水片断带运动解析数据(运动参数),数据量约 1.8G。

(2)舞台艺术类视频数据库

舞台艺术类视频数据库目前主要建立了群体动作部分,包括 43 个完整的群体舞蹈,约 1200 个片断,数据格式为 MPEG-1,数据量为 3G。

(3)表情动作视频数据库

表情动作视频数据库包括了 70 个人(男性 51 人,女性 19 人)的 1000 段脸部表情视频(说话类表情 500 段,情感类表情 500 段);数据格式采用 MPEG-2,分辨率为:640 × 480 的 24 位彩色,30 帧/秒,数据量约 3G。

2.5 人体生物特征类

人体生物特征类数据库目前建立了面像识别用人脸数据库,主要用于面像识别的关键技术研究和产品开发。数据库共有 1042 人,平均每人 100 张图像。数据库基本涵盖光照、姿态、表情、饰物、背景、时间、距离、年龄等主要与面像识别相关的因素,并有相应的组合。数据采用 bmp 格式存储,每张图像为 640 × 480 或 320 × 240,数据量为 80G。

2.6 多模式类

多模式数据库是为研究多模式人机交互技术而建立的,目前包括 2 个库:

(1)汉语听觉视觉双模态语料库

汉语听觉视觉双模态语料库,包括通用领域和

万方数据

奥运领域 2 个子库，采用 MPEG-1 格式存储，图像大小为 320 × 240，每秒 30 帧。其中：

库名	说话人数	语料	发音次数	数据量
通用领域双模态库	20 人	通用领域的 120 个单字、398 个词组和 111 个句子	3 遍	28G
奥运领域双模态库	20 人	向奥运领域的 500 个专业词汇和 100 个句子	3 遍	18G

(2)手语数据库

连续手语数据库主要用于手语识别与合成技术的研究与产品开发。目前的数据库是以《中国手语辞典》作为建库的标准，采用数据手套方式，由专业手语老师表演，建立了手语数据库，具体内容如下：

语料	人数	每人采集次数	合计样本数
2400 个词根	6 位手语老师	4	57600
5500 个词汇	8 位手语老师	2	44000
2000 个连续手语句子	8 位手语老师	2	32000

3 基础数据库的应用

为了推动基础数据库的应用，由中国科学院自动化研究所、中国科学院计算技术研究所、清华大学、教育部语言文字应用研究所等单位发起，成立了中文语言数据联盟[5]，即 Chinese LDC，隶属于中国中文信息学会，挂靠中国科学院自动化研究所。目前 Chinese LDC 的数据主要来源于国家重点基础研究发展规划(973)项目‘图像、语音、自然语言理解与知识发掘’和国家高科技研究发展计划(863)重点课题‘中文平台总体技术研究与基础数据库建设’中语言和语音相关部分的数据。基础数据库的共享应用，对充分发挥大规模基础数据库的共享优势，避免小规模数据封闭性重复建设，推进本领域关键技术的研究开发，加速我国信息化的进程和相关产业的发展具有极其重要的意义。

目前，863 基础数据库已经得到了较好的应用，其中：大字符集联机汉字识别样本库已用于汉王公司的识别系统中；语音识别数据库已在中科院自动化所、清华大学等单位使用，语音合成数据库已在清华大学和安徽中科大讯飞公司等单位使用；文本类语料库由华建和 TRS 公司用于奥运项目的关键技术研究中，人体运动数据库通过中科院计算所在国家跳水队和蹦床项目训练中应用；上海银晨网讯公司运用人脸数据库取得的成果获得了 2003 年度上海市科技进步一等奖，中科院计算所运用手语数据库取得的成果获得了 2003 年度国家科技进步二等奖。此外，基础数据库还在 2003 年度和 2004 年度的 863 计划中文信息处理与智能人机接口关键技术评测中发挥了重要作用。

4 结束语

从世界范围看，基础数据库的建设向着规模、深度和自然的方向发展。在基础数据库的规模上，要求达到足够的规模；在基础数据库的深度上，力求进行深层次的加工与标注；在基础数据库的自然方面，强调基础数据库的内容要尽量真实、客观。

参考文献

[1] The Linguistic Data Consortium (LDC) 网站：http://www.ldc.upenn.edu
[2] The Text REtrieval Conference (TREC 会议)：http://trec.nist.gov/
[3] NII Test Collection for IR Systems (NTCIR 会议)：http://research.nii.ac.jp/ntcir/workshop/
[4] Qian YueLiang, LIN ShouXun, Zhang YongDong, LIU Yang, LIU Hong and LIU Qun. An Introduction to Corpora Resources of 863 Program for Chinese Language Processing and Intelligent Human-Machine Interaction. In Proceedings of The 4th Workshop on Asian Language Resources (ALR-04), March 25, 2004, Sanya City, Hainan Island, China.
[5] 中文语言资源联盟(ChineseLDC)网站：http://www.chineseldc.org

Design and Construction of HTRDP Corpora Resources for Chinese Language Processing and Intelligent Human-Machine Interaction

Qian YueLiang, Lin ShouXun, Liu Qun, Liu Yang, Liu Hong, Xie Ying
(Institute of Computing Technology of Chinese Academy of Sciences, Beijing 100080)

Abstract

This paper briefly presents the significance and the state of the art of Corpora Resources for Chinese Language Processing and Intelligent Human-Machine Interaction, then focuses on the design and construction of HTRDP (High-Tech Research and Development Program of China) Corpora Resources. Finally, the application and future work are addressed.

Key words: chinese language processing, intelligent human-machine interaction, corpora resources

万方数据