国家自然科学基金在自然语言处理领域近年来资助的已结题项目综述*

徐 琳[1+], 赵铁军[2]

[1](国家自然科学基金委员会 信息科学部,北京 100085)

[2](哈尔滨工业大学 计算机学院,黑龙江 哈尔滨 150001)

Summarization of Results of Program Funded by NSFC in the Field of Natural Language Processing in Recent Years

XU Lin[1+], ZHAO Tie-Jun[2]

[1](Information Science Department, National Natural Science Foundation of China, Beijing 100085, China)

[2](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: +86-10-62327141, E-mail: xulin@mail.nsfc.gov.cn, http://www.nsfc.gov.cn

Received 2005-09-20; Accepted 2005-09-21

Xu L, Zhao TJ. Summarization of results of program funded by NSFC in the field of natural language processing in recent years. *Journal of Software*, 2005,16(10):1853−1858. DOI: 10.1360/jos161853

Abstract: In this paper, summarization of results of program funded by NSFC in the field of natural language processing in recent years is given, including summarization of Chinese information processing technology, natural language processing application technology, minority language information processing technology.

Key words: National Natural Science Foundation; natural language processing; Chinese information processing technology; natural language processing application technology; minority language information processing technology

摘 要: 对国家自然科学基金近年来在自然语言处理领域资助的已结题项目进行了综述,内容涉及中文信息处理技术项目总结、自然语言处理应用技术项目总结以及少数民族语言信息处理技术项目总结.

关键词: 国家自然科学基金;自然语言处理;中文信息处理技术;自然语言处理应用技术;少数民族语言信息处理技术;综述

国家自然科学基金近年来十分重视中文和我国少数民族语言的信息处理研究,在资助方面作出了一定的资金保证,期望大力推进我国自然语言处理研究领域的技术发展.针对目前已经结题的项目(见表 1),分 3 个方面进行概要介绍.

* 作者简介: 徐琳(1964－),女,浙江宁波人,博士,副教授,主要研究领域为计算机图形学,人工智能;赵铁军(1962－),男,博士,教授,主要研究领域为自然语言处理,机器翻译技术.

1 中文信息处理技术

国家自然科学基金对中文的词汇、句法、篇章分析方面的研究都给予了资助.鉴于中文本身的特点,这些受资助的项目在词汇、句子、篇章的语义侧面上都展开了研究,包括词汇语义计算、句法语义模型以及篇章语义的应用.

词汇是自然语言的基石,是语言更高层面自动分析的基础.“构建基于情境的词汇语义学的计算平台”项目从“情境表现意义的关系”的基本定义出发,尝试对词汇语义所表达的概念进行系统的整理和组织.该研究提出:概念的表达可以通过概念本身的生成过程来描写,而这一过程是和代表概念的具体词汇所处的上下文相关的,这样一个上下文环境被认为是情境[1].因此,该研究的核心内容之一是在概念的生成情境中定义概念和描述概念之间的关系,为此提炼了一套数学描述工具,进而形成完整的基于情境的词汇语义学描述体系;在人工提炼 140 个情境描述的基础上,总结形成了基于不同语义词典描述资源的人工提炼情境内容的基本准则和处理方法,应用在知网和同义词词林的信息融合方面的工作[2],为开发一个方便、灵活的汉语情境开发支撑平台提供了有力支持.该研究还在已有的句法研究(包括树库资源)的基础上探索了中文句法成分到语义资源的自动联结算法,以期形成汉语“句法→语义→概念”分析理解的联系通道.

词汇的搭配是描述词间组合能力的一种重要的词汇知识,在语句分析、词义消歧和信息抽取方面有着重要作用.“汉语动词搭配知识的自动发现研究”项目在中文大规模真实语料库(6 个月《人民日报》分词和词性标注语料库)基础上,进行了动词搭配知识抽取的实验,获取了 50 万动词搭配词对、1 万条动词语法搭配等知识.要获取高精度的词语搭配知识,势必需要对真实句子进行分析,以确定什么是动词的主语、宾语、修饰语等,进而获得有关的搭配.因此,该项目实施的基础包括完成了不同层次的部分句法分析,特别是建成了 200 万汉字的功能语块库,为课题的顺利完成提供了有力保证.

词汇级统计语言模型在自然语言处理领域有着广泛应用,国家自然科学基金资助了相关研究,“基于大规模语料库的汉语词语自动聚类研究”项目是其中之一.该项目以基于大规模语料库的汉语字、词的不同元数,尤其是三元以上的同现概率统计为基础,研究了汉语词语自动聚类的关键技术[3,4],包括汉语构词统计规律、基于上下文的词语相似度的计算方法、面向大词表的词语自动聚类算法,进而构造一个基于类的统计语言模型.

在句子语义表示的探索方面,国家自然科学基金资助了“面向自然语言处理的逻辑语义表达与演算模型研究”项目.该项目以词汇功能语法(lexical functional grammar,简称 LFG)为基本参考语法框架.研究设计了面向中文和中英文对比的自然语言语法语义一体化处理模型框架,即 LogSem 框架.该框架共有 3 个层次,即表层句法成分结构(sc-结构)、表层句法功能结构(sf-结构)、深层句法结构(ds-结构),其中 ds-结构可向自然语言语句的逻辑表达投射.目前,正在用 LogSem 框架作基础并结合本体论技术,进行相关机器翻译实验研究[5].使用细致结构信息的句法-语义分析技术能否经得住真实文本的考验,一直是自然语言处理研究面临的挑战之一.

篇章语义分析方面的典型应用是指代消解,“汉语指代消解与多文本交叉共指研究”项目对单个文本和多个文本之中的指代消解都作了探讨和实验.该项目针对单个文本,研究了人称代词与先行语的关系、命名实体的共指问题、指示代词消解所需的基本知识等.对于多个文本,研究了命名实体的共指关系,如相同人名在不同文本中出现时,是否表示同一对象的问题.相关的实验表明,人名、地名、机构名三大命名实体共指消解具有较高的准确度,其中单数第三人称“他”的消解,在人民日报语料库上测试,超过了 90%的准确度.在多文本命名实体的共指消解方面,选择了歧义现象最突出的人名同指消解,项目中所涉及的实验结果的 F-measure 值超过了 85%.该项目研究者对于利用弱化的语言知识,如单复数、性别特征等进行指代消解也作了深入研究[6].

2 自然语言处理应用技术

国家自然科学基金对自然语言处理应用技术给予了积极支持,期望相关技术在实际应用中能取得良好的效果.这些相关技术主要包括机器翻译、信息检索、自动文摘技术等.

机器翻译作为一个需要长期研究的挑战性课题,一直受到国家自然科学基金的关注,不断资助研究者从事该方向的研究与开发.“基于语段处理的网上英汉机器翻译系统”项目以网络大规模文本出现的新环境为背景,

研究了 4 种类型的基于实例的机器翻译方法(EBMT),即翻译记忆、基于词表的 EBMT、基于模板的 EBMT 和基于结构的 EBMT.在这些方法的研究中,核心内容之一是扩展语段(extended chunk)的定义.这是一种新的知识表示方式,形式上是一个无翻译歧义的单词或单词串[7].它是基于语义定义的,具有无歧义性、复现性、可嵌套性、内部结构句法自足性等特征.通过英汉机器翻译系统中的 3 类知识库:电子词典、E-Chunk 库和规则库,从而实现了上述基于实例的机器翻译方法.研究成果最终集成在一个以语音作为输入和输出的面向网上真实文本的口语英汉机器翻译系统之中.

对翻译机制的探索,也是推进机器翻译研究进步的一项内容.“基于知道逻辑的网上自动选择翻译方法研究”项目采取了蒙塔鸠语言理论和技术来进行源语言的理解和目标语言的生成研究,构造了一种能够进行信念修正的全新认知动态逻辑,从而为网上选择翻译机制的实现提供了坚实的逻辑推理系统.在汉语的机器理解方面,提出了一种意群动力学的新理论,并应用于词切分、标注以及指代消解等多个方面.这些研究也为认知逻辑本身的发展做出了贡献.

高效率的信息检索已经成为互联网时代自然语言处理的核心技术,检索形式的多样化、智能化,检索内容贴近用户需求是研究开发者共同追求的目标.“基于信息抽取和模板生成的多语种信息检索模型的研究”项目以特定信息,如命名实体的识别和相关的模板生成作为检索手段,以期提高检索精度.围绕着信息抽取和模板生成,该项目进行了汉语特定领域命名实体的自动识别和事件信息的自动抽取;提出了语言无关的模板与语言有关的模式的结合、事先定义的模板与动态获取的模板,克服了信息抽取面向领域的局限,实现多语种信息处理的一致性.同时研究了基于 Web 多语种词汇的在线自动获取技术,提高了多语种信息检索与抽取的查全率和查准率.提出了一种多语种信息检索的模型,实现了投资信息的多语种信息抽取实验系统[8].

以问答方式来实现开放领域的信息检索,能够满足用户的多方需求,是一种高效而合理的检索方法.目前问答式检索技术已经成为一个研究热点,属于自然语言处理和信息检索两大研究方向的结合部,国家自然科学基金也给予了重点资助.作为先期的探索[9],“开放域问答式信息检索技术研究”项目探讨了多种相关技术,包括:自然语言问句的分类和理解,答案的抽取和融合,用于问句扩展和答案验证的复述技术,支撑问题理解的词义消歧和句法分析技术,应对简述型问题的多文档自动文摘,对问答系统的自动评价等.在问答系统中,核心问题是“复述”.围绕着复述研究,建立了基于多译本的复述语料库,进行了复述语料句对齐、词对齐、复述转换模板的研究.此外,该项目还建立了为研究者共享的标准问答测试集,实现了相关实验系统,并开发了“金山在线客服问答系统”[10].

面对互联网浩瀚的信息,自动文摘是另一项受到普遍关注的技术,特别是多文档的自动文摘技术更成为近来的研究热点.“基于逻辑框架的多文档自动文摘技术”项目以网络大规模真实文本作为处理对象,探索了汉语多文档自动文摘模型的实现[11].该项目通过对网络文本聚类技术(基于概念的文档主题聚类算法)、信息融合技术(基于逻辑回归模型的文本片断聚类算法)、信息压缩技术(基于词汇链技术以及 MMR 方法的交叉文本结构压缩算法)以及文摘生成技术等文本级自然语言处理的深入研究,建立了一个汉语多文档自动文摘系统.该系统在网络文本聚类、多文档自动摘要的主题一致性以及摘要的可读性等方面取得了较好的结果.

3 少数民族语言信息处理技术

国家自然科学基金近年来十分重视少数民族语言信息处理技术的发展,为促进我国少数民族信息化建设做出了积极贡献.基金委专门设立以资助少数民族地区科研为主的地区科学基金就是一项重要措施.少数民族语言文字的自动化处理在国家自然科学基金资助的自然语言处理领域项目中占了相当大的比重,已经结题的项目中几乎有一半是此类项目.

少数民族语言信息处理的资助项目主要集中在蒙文、藏文和新疆地区民族语言上.下面分别加以介绍.

任何语言的自动处理都首先都要解决计算机输入和输出问题,少数民族语言因为用户数量少,民族文字的键盘输入和字形显示的有关研发工作相对滞后.国家自然科学基金对于此类研究也给予了很多支持.这也是民族语言信息化平台的基础部分.近年来,有 2 项关于蒙古文输入和输出技术、1 项关于藏文输出技术的资助.“蒙古文整词输入法重码词智能化选择输出技术研究”项目是在《蒙古文整词输入法》的基础上进一步进行研发,

重点解决重码词的选择问题.该项目主要研究了重码词获取方法、重码词搭配关系、重码词知识表示方法、重码词排歧选择方法和重码词选择的实现技术.该项目取得了预期成果,测试表明,该技术使平均每输入 10 个整词时出现一次重码词,降低为平均每输入 25 个整词时出现一次重码词.其开发的软件模块已被集成在原输入法中,构成了新的整词智能化输入输出软件系统,已进入试用和推广应用阶段.

“蒙文 TrueType 字型字体技术的研究”项目完成了蒙文 TrueType 字型的研制,可以在 Windows 环境中,应用蒙文 TrueType 字型字体输入、显示、输出或打印蒙文.在研制过程中,根据蒙文特征、特点、语法规律、正字法要求,考虑了美观性,注重体现蒙文字形的艺术加工,设计制作了蒙文字型字体样本原稿;按照相应的数学算法、笔形特点和结构特点,修正笔划样本、连体样本、印刷体和手写体、艺术造型体样本.确定了标准字体特征样本,并根据字型字体的位置、比例、高度、宽度、艺术造型规则,进行了测试分析研究工作.

“藏文字型生成与识别”项目把藏文的字型生成和印刷体识别结合起来,完成了现代藏文印刷体识别系统.该项目完成 9 种字体的藏文 TrueType 字库;参考其他软件的字体样张,设计生成了现代藏文字体库 1 730 个、梵文样本库 105 个.同时通过合作,完成了实用化“多字体印刷藏文(混排汉英)文档识别系统”,实现了藏文文档图像输入及版面分析,印刷藏文与汉英混排文本识别,识别后文本编辑(包括对照、候选字选择、插入、删除)等功能.测试表明,藏文白体、黑体等 6 种字体单字平均识别率达到 99.83%,实际藏汉英混排文本的平均识别率达到 97.28%以上.

民族语言信息处理平台建设,在多民族聚居地区具有重要意义.国家自然科学基金资助了“基于 LINUX 的新疆维哈柯汉英多语种信息处理平台”项目.该项目突破了 Windows 系统的限制,采用开放源码的 Linux 系统作为多民族语言文字的处理平台,以促进我国少数民族自主产权信息系统的推广应用及产业化[12].该项目的主要研究成果包括:掌握了在中文 Linux 系统平台中实现多语言输入输出、存储等技术;实现了基于 Linux 系统平台的具有不同编辑方向、不等宽、不等长代码、连笔书写特征的多语种混合处理技术;在操作系统核心上实现了支持 ISO10646(GB13000-1993)标准的多文种信息处理平台,提供对 GB18030-2000 编码集和维哈柯文国际代码的代码转换页;实现了维汉英多语种 Linux 信息处理系统的桌面软件和服务器版本.

自然语言处理离不开词典和语料库等资源.在少数民族语言信息处理的资源建设方面,国家自然科学基金也给予了资助.“蒙古语语法信息词典框架设计”项目设计了面向信息处理的、通用的蒙古语语法信息电子词典框架;从信息处理角度对蒙古语词语进行了分类,制定了蒙古语词类标记集;通过对蒙古语词语各种语法属性的研究,设计了易于机器处理的各种属性字段,并指定其取值规范;整个蒙语语法信息词典以数据库形式提供,采用了基于 ISO 10646 的蒙古文编码国际标准,包括 1 个总库和 15 个分库.“人机互助的通用现代维语语料库加工处理系统的研究”项目建立了人机互助的通用现代维语语料库加工处理模型.该模型的组成包括:收集原始语料及文本格式转换,标注语料所需的各类语言资源,如语法信息词典、规则库、统计信息库等,语料库管理和查询模块、统计处理模块、自动词类标注和人工校对模块、规则学习模块等各类处理模块.

各种民族语言信息处理的应用技术也在国家自然科学基金的资助下取得了一些研发成果.“基于词典和格标记的现代藏语自动分词系统研究”项目实现了以句法结构分析和形式标记识别为核心的组块识别及块内分词.自动分词离不开词典,因此该研究包括构建带句法语义信息的词典部分.同时对藏语的词法、句法、组块分析等进行了标注集建设、标记分类和识别,建立了面向信息处理的藏语语法体系.已建成的词典数据库包括 8 万余词条项,部分词项已填写各类词法和句法信息.与词典匹配的文本语料精选了约 3000 句式,建成了典型词语句法数据库.

“蒙文上网及蒙文全文检索的理论与技术研究”项目是针对蒙文的特殊性而展开的.由于各种流行的网页编辑器及 HTML 均不支持蒙文的“由左到右、竖行”书写格式,严重阻碍了蒙文信息上网.该项目重点研究了蒙文编辑控件、蒙文网页编辑和蒙文全文检索的关键技术、模型和实现方法,设计和实现了蒙文编辑控件(ActiveX 控件)进行蒙文编辑,在蒙文网页编辑器中对蒙文进行“所见即所得”的实地编辑,在浏览器中浏览蒙文信息.应用 DHTML 技术设计和实现了蒙文网页编辑器,它既是普通的网页编辑器,也是蒙文编辑控件的容器,使网页编辑和蒙文编辑融为一体.同时,还针对蒙文的特点,建立了蒙文全文检索模型和蒙文信息分类模型.

“新疆民文校对研究”项目针对新疆地区多种民族文字的自动文本校对提出了相应的解决方案.新疆民文

(包括维吾尔文、哈萨克文、柯尔克孜文)属于突厥语言的阿勒太语系,是一种自右向左、自上而下书写的拼音文字,其校对有其自身的特点.该项目主要以维吾尔文为主,以其他民族语言为辅,通过统计大量的语料,邀请民文语言学家参与,归纳出民文文本错误的种类,分为文本错误、数字错误、标点符号错误及其他错误.文本错误又分为非词错误和真词错误.针对总结出的维文文本错误的类型,采用基于民文规则和语料库统计的方法,机器自动查错与人工确认和查纠相结合,实现了先查错后纠错.在 Microsoft Word 2000/XP 软件上挂接了维文校对系统,处理单词级的拼写校对.纠错以非词错误为主,用模糊匹配算法和最小编辑距离对词进行排序选词,供人工纠错使用.

表1 国家自然科学基金在自然语言处理领域资助的已结题项目列表

项目编号	项目名称	负责人	执行期限	资助金额(万元)	项目类型
69573026	彝文信息处理及标准彝文字符集与编码的研制	沙马拉毅	1996.1-1998.12	6	自由申请项目
69903007	汉语动词搭配知识的自动发现研究	周强	2000.1-2002.12	12	青年基金项目
69963001	蒙古语语法信息词典框架设计	那顺乌日图	2000.1-2002.12	10	地区科学基金项目
69963002	新疆民文校对研究	古丽拉	2000.01-2003.12	10	地区科学基金项目
69973015	基于大规模语料库的汉语词语自动聚类研究	王晓龙	2000.01-2002.12	12	自由申请项目
69983006	基于知道逻辑的网上自动选择翻译方法研究	周昌乐	2000.1-2002.12	11	高技术探索
60063001	蒙文上网及蒙文全文检索的理论与技术研究	王俊义	2001.01-2003.12	14.5	地区科学基金项目
60063002	蒙文TrueType 字型字体技术的研究	巴力登	2001.01-2003.12	14.5	地区科学基金项目
60073058	藏文字型生成与识别	于洪志	2001.01-2003.12	16	自由申请项目
60083003	基于信息抽取和模板生成的多语种信息检索模型的研究	盛焕烨	2001.01-2003.12	13	高技术探索
60083005	词汇、句法和语义——基于认知实验的汉语加工过程研究	孙茂松	2001.01-2003.12	13	高技术探索
60083006	基于语段处理的网上英汉机器翻译系统	姚天顺	2001.01-2003.12	13	高技术探索
60103014	基于内容的文本过滤技术研究	黄萱菁	2002.01-2004.12	28	联合资助基金项目
60163001	基于LINUX 的新疆维哈柯汉英多语种信息处理平台	吾守尔·斯拉	2002.01-2004.12	17	地区科学基金项目
60163002	人机互助的通用现代维语语料库加工处理系统的研究	玉素甫	2002.01-2004.12	17	地区科学基金项目
60173005	汉语指代消解与多文本交叉共指研究	王厚峰	2002.01-2004.12	18	自由申请项目
60173008	构建基于情境的词汇语义学的计算平台	周强	2002.01-2004.12	18	自由申请项目
60173024	基于词典和格标记的现代藏语自动分词系统研究	江荻	2002.01-2004.12	19	自由申请项目
60173025	面向自然语言处理的逻辑语义表达与演算模型研究	王惠临	2002.01-2004.12	18	自由申请项目
60203020	开放域问答式信息检索技术研究	刘挺	2003.01-2004.12	15	青年基金项目
60263003	蒙古文整词输入法重码词智能化选择输出技术研究	S·苏雅拉图	2003.01-2004.12	7	地区科学基金项目
60373100	基于逻辑框架的多文档自动文摘技术	王晓龙	2004.01-2004.12	8	自由申请项目

References:

[1] Chen ZS, Zhou Q, Zhao Q. Situation——A suitable framework for organizing and positioning lexical semantic knowledge. Computational Linguistics and Chinese Language Processing, 2002,7(2):1−36.

[2] Mei LJ, Zhou Q. Merge information in hownet and TongYiCi CiLin. Journal of Chinese Information Processing, 2005,19(1):63−70 (in Chinese with English abstract).

[3] Liu BQ, Wang XL. User-Oriented Chinese language model and its machine learning. Journal of Harbin Institute of Technology, 2004,36(2):150−153 (in Chinese with English abstract).

[4] Zhao Y, Wang XL, Liu BQ, Guan Y. Applying class triggers in Chinese POS tagging based on maximum entropy model. In: Proc. of the 3rd Int'l Conf. on Machine Learning and Cybernetics. Vol 3. 2004. 1641−1645.

[5] Yang HY, Wang HL. A study of MT oriented ontology building. 2005. http://preprint.nstl.gov.cn/newprint/Upload/2005/1120115765651.pdf

[6] Wang HF, Mei Z. Robust pronominal resolution within Chinese text. Journal of Software, 2005,16(5):700−707 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/700.htm

[7] Lü XQ, Guo J, Yao TS. Knowledge bases in English-Chinese machine translation system ECT. Mini-Micro Systems, 2004,25(8):1482−1485 (in Chinese with English abstract).

[8] Li F, Sheng HY. Research on information retrieval and information extraction. Application Research of Computers, 2002,19(1):16−18 (in Chinese with English abstract).

[9] Zheng SF, Liu T, Qin B, Li S. Overview of question-answering. Journal of Chinese Information Processing, 2002,16(6):46−52 (in Chinese with English abstract).

[10] Zhang Y, Liu T. Experimental investigation of online client-server based on set of familiar question. In: Proc. of the 8th National Computational Linguistics Joint Conf. Beijing: Tsinghua University Press, 2005. 474−480 (in Chinese with English abstract).

[11] Chen YM, Wang XL, Liu YC, Lou XZ. Automatic text summarization based on topic and content. Computer Engineering and Applications, 2004,40(33):11−15 (in Chinese with English abstract).

[12] Mu C, Yuan BS, Silamu W, Li LW. Design and implementation of processing platform for Uighur, Kazak, Kirgiz, Chinese and English. Computer Engineering, 2004,30(10):71−73 (in Chinese with English abstract).

附中文参考文献:

[2] 梅立军,周强.知网与同义词词林的信息融合研究.中文信息学报,2005,19(1):63−70.

[3] 刘秉权,王晓龙.一种面向用户的语言模型及其机器学习的方法.哈尔滨工业大学学报,2004,36(2):150−153.

[5] 杨海燕,王惠临.面向机器翻译的本体构建初探. http://prep.istic.ac.cn/eprint/Upload/2005/1120115765651.pdf

[6] 王厚峰,梅铮.鲁棒性的汉语人称代词消解.软件学报,2005,16(5):700−707. http://www.jos.org.cn/1000-9825/16/700.htm

[7] 吕学强,郭军,姚天顺.英汉机器翻译系统 ECT 中的知识库.小型微型计算机系统,2004,25(8):1482−1485.

[8] 李芳,盛焕烨.信息检索与信息抽取技术的研究.计算机应用研究,2002(1):16−18.

[9] 郑实福,刘挺,秦兵,李生.自动问答综述.中文信息学报,2002,16(6):46−52.

[10] 张宇,刘挺.基于常见问题集的在线客服实验研究.见:第 8 届全国计算语言学联合学术会议论文集.北京:清华大学出版社,2005.474−480.

[11] 陈燕敏,王晓龙,刘远超,楼喜中.一种基于文章主题和内容的自动摘要方法.计算机工程与应用,2004,40(44):11−15.

[12] 缪成,袁保社,吾守尔·斯拉木,李莉维.哈、柯、汉、英多文种处理平台的设计与实现.计算机工程,2004,30(10):71−73.