中 文 信 息 学 报
第 13 卷 第 4 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 13 No. 4

蒙古文自动处理系统研究

嘎日迪 赵小兵 马红旭* 赛 音* 白小玲*

内蒙古师范大学 010022 *内蒙古自治区电子计算中心

摘要 本文主要介绍了蒙古文自动处理系统的构成;蒙古文语料库,蒙古文知识库,蒙古文数据库以及蒙古文自动处理系统的软件等问题的初步探索研究的过程。

关键词 蒙古文自动处理 蒙古文语料库 蒙古文知识库 蒙古文数据库

Study on Automatic Processing System of Mongolian Language

Garidi Zhao Xiaobing Ma Hongxu* Sai yin* Bai Xiaoling*

Inner Mongolia Noraml University 010022

* Inner Mongolia Electronic Computing Centre

Abstract This paper mainly introduce how to constitute Mongolian automatic processing system. It includes the Mongolia phraseology library, the Mongolian linguistry library, the Mongolian data library, and pilot study course of software on the Mongolian automatic processing system.

Keywords Mongolian automatic processing Mongolian phraseology library Mongolian linguistry library Mangolian data library

早在 80 年代初,北京语言学院的张普教授就提出了语言自动处理系统的论点,也就是语言现代化的概念。他认为:广义地讲,利用计算机对语言信息进行自动加工处理的工作都属于语言自动处理的范畴。具体地说:语言自动处理系统是包括利用计算机按照语言内部的各个结构系统(语音、文字、词汇、语法、标点和外文字符号等系统),对一段特定的语言进行分类、分析、统计、归纳、综合、比较,以帮助人们更深刻、更科学地认识和研究应用语言的内部规律的一门学科。语言自动处理是一门集语言学、计算机科学、数学、心理学和自动化技术等学科在内的综合性的边缘学科。它的研究对象包括语言规律和模式及计算机软件两大方面。

根据张普教授提出的论点,我们研制了一个蒙古文自动处理系统,并进行了探索性的研究。现将有关情况进行交流。系统由蒙古文语料库、蒙古文知识库、蒙古文数据库及自动处理软件四大部分组成,在此进行逐一介绍。

本文于 1999 年 1 月 18 日收到

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一、蒙古文语料库

语料库的内容从逻辑上讲是越多越好,而且应包含方方面面的内容。但因环境、资源、时间,精力以及学科等诸多因素决定,语料库的资源不可能无限。所以,只能采用统计学的方法从众多的学科资料中科学地选择和抽样选取,并且使语料的数量达到了一定的数量级。我们选择了 150 万字的量级。

语料抽样选择的总体质量是否恰当,直接影响结果的代表性、客观性和正确性。我们挑选语料的基本着眼点是应针对各种类型的蒙语教学、科研及应用最广泛的一般需要,满足具有相当于我国蒙古族中等文化水平的人在日常工作和社会生活中使用蒙古语词的需要,并为蒙古语的语言规范化和蒙古文信息处理的标准以及为语言教学和其它相关学科的研究提供有科学价值的参考材料和数据。

我们在实际工作中参考了汉字的经验,同时根据蒙古文的实际使用情况,在选材比例上有所侧重。确定以日常口语、科普书刊、报刊政论、文学作品等四类题材作为选取语料的来源。

各种题材的语言作品有四类,共 170 种,其中:

第一类:戏剧、民间故事、数来宝、相声、广播稿件、散记等日常口语材料。选用反映日常生活各个侧面的剧本、名著。另外,考虑到日常交往中口语的实际使用情况,还先用了专题采录和随机采录的部分口语材料,共 31 种,约 14 万词以上,占全部语料的 10 %。

第二类:科普书刊材料。选用中等文化水平范围内的科普文章。选用生物、医学、工程、技术以及与衣食住行有关的通俗科技文章 32 种,约 28 万词以上,占全部语料库的 20 %。

第三类:报刊、政论文章及名著。选用反映我国现实社会生活的政治、经济、法律、历史、地理、军事等方面的材料 33 种,约 42 万词以上,占全部语料总量的 30 %。

第四类:各种题材的文学作品。选用长篇,中篇,短篇小说、散文、童话、传记等,共 84 种,约 56 万词 ,占全部语料总量的 40 %。

语料选材基于以下几个方面的考虑:

(1) 以现代蒙古文优秀文学作品为主,侧重在语言运用上较为成功的作品,兼顾各种不同的语言风格及文学流派。

(2) 较多选用 50、60、70 年代的作品。在题材上尽量照顾到各方面的内容和描写对象,力求内容大体平衡。

(3) 保持原作的完整性。短篇尽量全文选用,中、长篇著作则按原作章节截取相对完整的并具代表性的片段。

就全部的语料来说,为了使包括的词汇全面和客观,对部分语料还采取了等距抽样的方式选取材料。

选取的材料不能数量太少,否则不能说明问题。然而,数量多虽然能加强精确性,但冗余度太大,不够经济。我们在选择语料时,考虑到选词的总体个数应足够多,同时也注意到了必须尽量保证所选语料具有代表性、多样性和均匀性。

当然,根据工作的需要还可在语料库中随时增加各种语料。还可按时期分类(例如中世纪语料)、著作分类、文体分类、其它分类和学科分类;其中学科分类又可详细分为自然科学分类和社会科学分类;而在社会科学分类中又可再次细分为文、史、哲、经济等等。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

二、蒙古语言知识库

语言知识库中存储着人们赋予计算机的各种语言文字信息知识。计算机应用这些知识,对语料库中的语料进行自动处理。例如,从语言、文字、词汇、语法等众多角度进行自动处理。各种处理过的数据结果,反过来又不断地完善、更新知识库的知识。

知识库的"知识",可分为语言知识、文字知识、语法知识、词汇知识和其它知识。这些知识需要存储在计算机中,供专用软件随时调用,此时往往要转化成各种语言模式。这些语言模式的建立是在传统的蒙古语言研究的基础上形成的,但又不同于传统的语言研究。它给蒙古语文工作者提出了更新、更严格的要求。从某种意义上讲,这种研究和知识主要是面向机器的,而不是面向人的。现代科学技术的发展使得语言超越了作为人类社会交际的工具这一范畴,开始了语言也能作为人和机器进行交流的时代,这就是通常所说的"人—机对话"。而对自然语言理解这一课题,也不仅仅是为了解决人对自然语言的理解问题,更主要的是解决机器的理解问题。在我们的语言知识库中做了以下几个蒙古语知识系统:

2.1 蒙古语复合词的确定和自动切分知识系统

蒙古语中大部分的词与词可用词与词间的空格就能判清。但对蒙古语的复合词就不能采用空格来判断了。若对蒙古语的复合词不能正确判断,就不能对语料库的语料进行正确的切分、统计和分析。

对蒙古文复合词如何划分,在理论和实践中都尚未得到妥善解决。但复合词在蒙古文中确确实实客观存在。例如,"人民"(ᠠᠷᠠᠳ ᠤᠯᠤᠰ),"中国"(ᠳᠤᠮᠳᠠᠳᠤ ᠤᠯᠤᠰ)等。因此,在蒙古语文知识库中必须有复合词的正确切分处理。但是,蒙古文至今还没有一个复合词词典。因此我们将内蒙古地区流行的几部蒙古文词典中出现的、无争议的蒙古文复合词全部收集起来,作为计算机切分蒙古文复合词的标准知识词典。利用这个标准知识词典来作为依据,对语料库中语料进行计算机的自动切分、统计和处理。我们收集到的蒙古文复合词约有三万三千多条。对于那些在理论上还有争议,暂不能确定的蒙古文复合词,待语言学界形成一致意见后,再收进蒙古文复合词词典中,使知识库不断地完善。

2.2 蒙古文同型多义词知识系统

蒙古文中有许多的同型多义词。对这些同型多义词,处理语料库时,必须按它的真正意义去分析和处理,否则会影响准确性。为此,我们做了蒙古文同型多义词知识词 典,并对其含义一一整理。(有的同型多义词有多达 5～6 个意思)。对语料中的同型多义词预先进行特殊处理标注,然后对语料进行分析、统计、处理,使自动处理的结果更加准确和科学。

我们建立的蒙古文同型多义词知识词典约收集了三千多条。随着研究工作的深入发展,还可不断地增加同型多义词的词条,使知识库不断扩大。

2.3 蒙古文附加成分知识系统

蒙古文的附加成分使用频率极高,而且词与附加成分之间用空格分开,不进行特殊处理,在统计工作中,就可能变成单词来统计。况且有的附加成分和某一单词同型,有时在切分统计时需要前面的单词一起进行切分统计,有时则不一起统计而单独统计处理等众多不同的要求。因此,我们做了一个蒙古文附加成分库,为这些特殊要求的完成奠定了知识基础。

2.4 蒙古文动词处理知识系统

蒙古文的动词变化情况在蒙古文应用中表现得十分活跃和丰富。在实际使用过程中如何变化,对动词如何进行统计和切分处理,需要了解清楚动词的时态变化等众多知识。为此,我

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

们做了蒙古文动词处理知识系统。

2.5 校对知识系统

在百万词以上的语料库的建设中,必须保证语料的正确性。这些语料是人工输入的,难免不发生错误,为了保证语料库中语料的正确性,必须对语料进行校对。人工校对费时、费精力。所以为了采用计算机对语料库的语料进行自动校对,我们做了正字法词典。把输入好的语料与计算机内部的正字法词典进行对照检查,修正语料库中的语料错误。

2.6 特殊要求知识系统

对语料库中的有些词汇,有时需要特殊处理。例如,语料中的人名、地方、数码和其它专用符号等需要特殊处理。因此,预先做了特殊处理知识系统。

三、蒙古文数据库系统

蒙古文数据库系统是应按数据结构原理,科学地存储在计算机硬盘上的各种自动处理过的蒙古文科学知识数据所组成的库。在库中尽量减少冗余度,供人们从各种角度灵活方便地检索和咨询。广泛地应用在蒙古文的学习、使用和研究中。数据库的建立是蒙古文自动处理系统中的最为困难的课题。

四、蒙古文的自动处理系统的软件构成

蒙古文自动处理软件,广义地讲,就是采用建立起来的知识系统来对语料库中的语料进行自动处理的各种软件系统。这些软件始终遵循快速、准确、方便的宗旨,最大限度地利用计算机资源进行自动处理,减少人工干预,采用自动的工程化管理。在需要人工处理的地方尽量提供方便的操作界面和帮助。能达到应用灵活自如的软件系统。具体地讲,包括蒙古文复合词的自动切分软件;蒙古文同型多义词的判断处理软件;蒙古文附加成分的特殊处理软件;蒙古文动词处理软件;蒙古文自动校对处理软件;语法分析软件;句法分析件等众多的软件系统。随着研究工作的进展,各种功能的软件不断增加,自动处理系统的功能也就得到了不断地完善与提高。

五、蒙古文自动处理系统和蒙古文现代化

蒙古文的自动处理系统为蒙古文研究提供了一种现代化的手段。利用这种手段,人们可以对各种蒙古语言实施科学的抽样,采用数理统计的方法实行"定量—定性"分析,从而得出科学的、准确的、全面的结论,避免了悬而不决。应当说自动处理的手段和统计法的真正应用是语言学家和语言研究的一项新的探索。

语言的频度特征是不可忽视的特征。传统的语言研究没有摆脱手工劳作的方式。基本上是采用"举例法"来论述问题。它的公式是"举例—定性"。但在实际的语言事实中偏偏有众多的例外。这就导致了另一公式的出现"举例外的—争论—不能定性"。于是形成了百家争鸣的众多悬而未决的问题。更何况蒙古文中有许多问题本来就处于模糊状态,所以单靠举例法是不能奏效的。事实上,许多人已经用手工进行各种统计研究。由于这种工作艰苦、繁琐而又枯燥,其统计结果也往往是局部、粗略、少量的。尽管如此,我们仍可认为这已经开始摆脱了"举例—定性"的公式,开始了语言的"定量—定性"分析,是一种真正的动态语言研究。在语言学

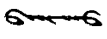
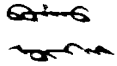
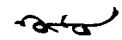
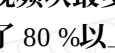
© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

领域里,已经发展出了一门新兴的学科 —计算机语言学或计量语言学。

忽略了蒙古文的频度特征,并不是传统蒙古语言学的过错。大量的统计工作、计算语言学的真正形成,均是借助于现代化的研究手段,单靠手工是无法进行的。因此,蒙古语言学的现代化包括三个主要组成部分,即理论的现代化、方法的现代化和手段的现代化。并且这三个组成部分互相促进、互相推动,又产生出了新的观点、新的结论。

下面是 150 万量级的蒙古文语料库进行自动处理的初步量化的有关数据:

1. 使用频率最高的十个词

序号	词 汇	蒙古文音标	出现的频次	频 率	累计频率
1		baihu	28224	2.8693	2.8693
2		gehtt	21759	2.2120	5.0813
3		bolhu	17149	1.7434	6.8246
4		ttgei	13384	1.3606	
5		yum	9376	0.9532	
6		mige	8775	0.8921	10.0305
7		httmttn	7962	0.8094	10.8933
8		cu(ctt)	7366	0.7488	11.5887
9		ene	7189	0.7303	12.3196
10		tere	6392	0.6498	12.9694

当出现频次最多的词汇为 268 个时,累计频率达到了 50 %以上;为 1715 个词 汇时,累计频率达到了 80 %以上;3897 个词汇时,累计频率达到了 90 %以上;6939 个词汇时,累计频率达到了 95 %以上。

2. 使用频率最高的独立形的五个附加成分

序号	形状	蒙古文音标	出现的频次	频率	累计频率
1		un	42212	3.14671	3.14671
2		yin	31279	2.3317	5.47841
3		u	30811	2.29682	7.77523
4		du(dtt)	26764	1.99513	9.77036
5		ni	24848	1.85280	11.62316

3. 使用频率最高的五个复合词

序号	形状	蒙古文音标	出现的频次	频率	累计频率
1		un	42212	3.14671	3.14671
2		yin	31279	2.3317	5.47841
3		u	30811	2.29682	7.77523
4		du(dtt)	26764	1.99513	9.77036
5		ni	24848	1.85280	11.62316

以上仅是以量化后的少量数据为例,更详细的数据如动词变化情况、语法树等内容,我们准备在其它文章中论述。

六、把民族语文(蒙古文)现代化课题纳入国家重点规划

上述已完成的工作应该说仅仅是蒙古文自动处理系统中较为简单的部分,而更艰巨、更有意义、更能吸引人的高难课题还等待着人们去开发和研究。例如,蒙古文与其它自然语言间的机器翻译;蒙古文字的自动识别技术;蒙古语的语音识别、语音合成;蒙古文的教学软件;蒙医

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

诊断治疗系统;懂得蒙古文的智能机器人等众多课题。这些课题之所以吸引人是因为它们是人类社会科学技术发展趋势的重要方面,它们的解决将会对蒙古族人民的生活产生巨大的影响,它们是代表社会现代化的显著标志。所以,这些课题应该纳入国家重点规划,给予一定的资金投入和人力投入。

参 考 文 献

[1] 张普. 论语言自动处理系统. 中国中文信息学会年会论文集,1983

[2] 陈原. 现代汉语定量分析. 上海:上海教育出版社,1989

[3] 黄昌宁. 关于大规模真实文本处理的谈话. 第三届中文信息处理国际会议论文集,1992,36

[4] 刘源. 关于多音字频度的统计. 第三届中文信息处理国际会议论文集,1992,234

[5] 梁南元. 再论汉语自动分词和切分知识. 中文信息处理国际会议论文集,1987,454

[6] 林联和. 谈汉语词汇统计文法. 中文信息处理国际会议论文集,1987,510

会议消息

全国第五届计算语言学联合学术会议(JSCL - 99)

1999 年 11 月 1 - 3 日 北京 龙泉宾馆

为促进国内计算语言学的研究和应用,加强同行间的学术交流与合作,中国中文信息学会、中国计算机学会、中国人工智能学会和北京市语言学会共同发起于 1999 年 11 月 1 - 3 日在北京龙泉宾馆联合举办“全国第五届计算语言学联合学术会议(JSCL - 99)”,会议的正式语言为中文与英文。

会议已发出录用通知,会议论文集将由清华大学出版社正式出版。

有意出席会议的请与清华大学陈群秀联系。

通讯地址:北京 100084 清华大学计算机系 陈群秀 收

联系电话:(010) 62787034

电子邮件:cqx @s1000e. cs. tsinghua. edu. cn

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net