

正确引导汉语理解与汉语研究

——事关人工智能开发的一个重要前提

Correct Guidance of Chinese Language Understanding and Chinese Language Research

林杏光

(中国人民大学对外语言文化学院,教授 北京 100081)

人工智能(利用计算机模拟人类智力活动的学科)的研究已成为当前最热门的研究学科之一,而自然语言理解(Natural Language Understanding,简称NLU)是人工智能的一个分支。自然语言理解有时也称计算语言学(Computational Linguistics),前者侧重于应用,后者侧重于理论,其内涵是一致的。汉语理解属于计算语言学范畴,是自然语言理解的一部分。汉语研究,在某种意义上说,是汉语理解的基础,因为汉语理解必须应用汉语研究的成果。

1974年经周总理亲自批准的“748”工程,标志着我国中文信息处理技术首开先河。由于方块汉字的特殊性,在汉语词句、篇章理解之前,首先要解决汉字处理阶段的问题。经过20多年的努力,特别是通过“七五”、“八五”的努力,中文信息处理技术已在汉字处理上取得相对的突破,以国标字符集为基础的汉字处理成果已获得广泛的应用。目前,对词语、句子和篇章理解的攻关已经开始,作为知识工程或智能化系统已经进入了实质性的知识获取或智能模拟的阶段。在这历史性的转变阶段开始之际,选择正确的汉语研究和汉语理解的路线,便显得至关重要。

从汉语“语汇最重要”, 看中外语言研究的异同

对语言的研究,中国和西方有相同的地方,也有不同的地方。相同的地方是都以古代书画文献为起点,以实用为目的。不同的地方是彼此发展的道路不相同。在西方,语言研究发端于语法学,词汇和语音、语义的研究长期作为语法学的附庸。而在中国,语言研究发端于词汇学,研究古典文献字(词)义的训诂学远比音韵学、文学学早,至于语法学更是远远地落在训诂学之后。总之,外国的语言研究

是语法研究率先,而中国的语言研究却是词汇研究率先。

张志公先生在80年代以后的几篇论文中(参见本文的“主要参考文献”)提出了“语汇最重要”的语言观。张先生认为:语言包括四个要素,即语汇、语音、文字、语法。这四个要素分为三个层面。第一个层面是语汇,它是概念、思维的物质基础。第二个层面是语音和文字,它们把语言交际的过程现实化,前者形于口语,后者成为书面语言。第三个层面是语法,它是组合语汇的规则。这三个层面,比较起来,语汇最重要,其次是语音和文字,再次才是语法。普通话说“你先走”,广州话说“你行先”,二者的语法规则不同,但都可听懂。如果没有“你”、“先”、“走”、“行”这些语汇,那就如同北京人口头常说的“没词了”,只能停止交际。俄语中的“书”(knuka)是阴性,德语中的“书”(das Buch)是中性,汉语中的“书”没有性,同样可以表达清楚。英语two books有了“二”还要加“s”,汉语就不用。由此可见,在语言里,最重要的是语汇,不是语法。张志公先生关于“语汇最重要”这一语言观正好和近10多年来世界语言学发展的八大倾向之首即“词汇主义”不谋而合。这说明在任何语言里,词汇都是最重要的,并不是仅在汉语中词汇才最重要。既然都是词汇最重要,为什么只是中国的语言研究以词汇研究为率先而外国的语言研究却以语法研究为率先呢?我认为这主要是汉语汉字和外国语言文字的不同特点造成的。

我国的春秋战国时代,“百花齐放”,“百家争鸣”。到了汉武帝的时候,罢黜百家,独尊儒术。之后的几千年,儒家的著作被奉为经典,研究它的学问被称为“经学”。由于语言文字的变化,记载经文的语言越来越脱离口语。先秦的经文,汉朝人便看不懂,非注释不可;到了唐宋,不但经书的原文看不懂,连汉朝人的注释也看不懂,对注释又要作注释。

这么一来,为了读经,就要开展语言文字的研究。首要的研究对象是直接影响读经的汉字。由于记录汉语的汉字是形、音、义结合的表意文字,字与词之间存在较一致的对应关系,使古人认为认识了汉字就是学会了汉语,研究汉字就是研究汉语。因此,几千年来,我国的语言研究主要集中在“字”的形、音、义上,相应地建立了文字学、音韵学、训诂学。而研究汉字不像研究西方语言的词那样会碰到属于语法学内容的词形变化的规则,因此,在马建中的《马氏文通》出版之前的一段相当长的时间里中国没有像现在大家所熟悉的系统的语法研究和语法学(词类,句子成分,单句,复句,等等)。外国的情况则不同。四大文明古国中的印度、希腊很早采用表音文字。这种文字与词明显与汉语不同,它只有形和音的问题,一涉及义就要联系到词。由于是形态语言,一研究词就碰到词形变化,一研究词形变化,便产生了语法的一部分重要内容即词法。这正是西方国家的语言文字研究发端于语法学的原因。

以上所述中外不同的语言研究情况表明,汉语和汉字是没有形态变化的,西方语言是有形态变化的。学习和研究有形态的语言首先碰到的拦路虎是语法,所以西方的语言研究须由语法学率先;而学习和研究无形态变化的汉语汉字首先碰到的拦路虎不是语法,所以中国的语言研究从最重要的词汇研究起步。

汉语“以意治形”与外语 “以形统意”的不同和对照

汉语有没有语法呢?回答是肯定的。但是,汉语语法学必须符合汉语的特点。汉语决定句子能否成立,很大程度上是意义的联系。汉语构词成句基本上是由意义之间的搭配关系决定的,而这些意义没有形态标记。印欧语有许多语言单位表示“数、格、时、态、性”等的形态,因此印欧语有明显的以形统意的特点,这和汉语以意治形的特点形成鲜明的对照。汉语的“话”如行云流水,句与句之间似断若连,它不注重形式,而注重内在的意念贯穿相承;不注重一个个孤立的句子,往往围绕一个话题展开议论叙述。在复句中间,没有关联词语标示的、靠意合组成的语句很多。拙著《复句与表达》专门写了一节“意合法是表达分句间逻辑关系的常用方式”。这表明,汉语是没有形态变化的语义型语言,以意合为主是其一大特点。汉语的语法学应着重研究汉语意合的方式和规律。

汉语不但无形态变化,而且在思维表达的方式上也有自己的特点。如汉语习惯的自然顺序是先整体,后局部;先共性,后个性,而印欧语言则恰恰相反,所以东西方的地址表达顺序正好颠倒。中国人

写信封从大单位到小单位最后到收信人姓名,外国人则相反,从收信人姓名到小单位最后到大单位。再如在时间表达的方式上,汉语和某些语言也正好相反。汉语说“吃完晚饭散步”,英语却是“散步于晚饭后”。这类例子可以举出很多。概括地说,汉语的语序和时间顺序一致,而像英语这类印欧语则不然,其语序往往和时间顺序相反。这些都从根本上确立了汉语语法学研究的重要出发点。

正当中国的语言研究还未及建立富有汉语特点的语法学的时候,1898年马建忠的《马氏文通》出版了,它引进了西方语言的语法模式。马氏的引进,功不可没,我们应该感谢马建忠为代表的老一辈汉语研究学者率先引进了西方的语法学体系。问题是引进之后,还应该进一步消化、改造,以至创造出适合汉语特点的语法学。由于未能很好地解决在引进的基础上进行创造的问题,使得汉语研究出现以西方语言学理论套在汉语上面的状况。所幸的是随着汉语研究的不断深入,已有愈来愈多的学者认识到,西方语言学理论总的来说是在形态语言的基础上建立起来的,用形态语言的理论去描写无形态的汉语,显然是不对路的。如此不对路的汉语研究成果当然也不能用以分析汉语的真实文本。这里谈点语文教学的历史往事可能是有意义的。1959~1979年,我在北京外国语学院(现易名“北京外国语大学”)向外语专业的本科生教授汉语语法,教学目的是使学生能应用汉语语法去指导翻译,以提高翻译水平。开始采用一般综合大学中文系的教法,即讲授语法理论,介绍不同流派的观点,诸如“池里养着一条鱼”的“池里”,一派认为是主语,一派认为是状语;“池里很干净”的“池里”,各派都认为是主语;“台上坐着主席团”的“主席团”,有人认为是宾语,有人认为是主语。老师自觉讲得津津乐道,但听课的学生不但不喜欢,而且还表现出厌烦。后改用分析真实文本的教法,即教师带领学生用语法知识去分析文章中的语法,学生很感兴趣,课堂很活跃。可是,若用外国语言语法模式套过来的汉语语法去分析汉语真实文本,就出现了可此可彼的问题,模棱两可的地方实在太多。50年代中期,教育部试图在中学的语文课上教语法,责成张志公编了“暂拟语法体系”,只试行了一年便难以为继。80年代初,教育部又试图在中学语文课上再教语法,责成张志公又编了本“语法试行提要”,结果还是教不下去,最后只能作出“淡化语法教学”的决议。总之,语文教学的情况表明,现在流行的汉语语法学不能用以指导语文教学实践。要分词,却没有词的定义;要标词性,名、动、形的界限又不清楚,词法和句子成分也不对应;要分析句法,主、宾语的界限,单句和复句的界限都不清楚。事情很显然,一个不能用于语文教学的东西能作为汉语理解的基础吗?因此,为了

汉语理解事业的顺利发展,我们应当克服中国文化传统中“求同易,求异难”这一消极倾向带来的困难,鼓足勇气改变现有的汉语研究路线。

当前在计算机汉语理解研究上,应该有新做法

当基于新的研究路线的汉语研究成果未能推出之前,汉语理解也不能坐等。但是面对目前的汉语研究状况和汉语汉字的特点,计算机汉语理解研究应该选择适宜的搞法。人之所以能理解语言是因为人具有语言知识,因此,要让计算机理解语言就必须为计算机建立语言知识库,这是大家的共识。但如何建立语言知识库?这是在国内外语言信息界都未取得共识的问题。我国语言信息界的大部分学者虽然对汉语理解还没有提出总体的理论框架和具体的实施方案,但实际上已经形成了初步的思路,那就是建立句法库、语义库、语用库及其他知识库。语用库现在显然还提不到研究的日程上。句法库包括分词、标注义项、标注词性、分析句法等功能。由于词类和句法成分是两个长期老大难问题,汉语语法学者已研究了100年仍未能得到解决,再研究下去看来还是无法解决,因此句法库的建立也碰到了简直无法逾越的障碍。那么语义库怎么建?大家倾向认为应该开发机器语义词典。但机器语义词典应该包括哪些内容?眼前也仍是一片朦胧。近几年来,大家觉得除语义分类系统外,还有两个内容应该描述:一是动词及其前后名词组成的框架;二是动词框架内部的语义制约关系,这个工作目前较倾向于采用格语法、原则参数语法和配价语法作为理论基础。但是,以上这种大部分学者倾向的思路能否解决汉语的理解问题呢?我仍深表忧虑。因为本文前面已经谈到,这种汉语理解的思路是建立在套用西方语言理论模式的汉语研究基础上的。另一种汉语理解的搞法是既提出了总体的理论框架,又设计了符号体系、知识库、理解处理模块等一系列具体的落实措施,而且还有一支由总设计者统一指挥的研制队伍,这就是中国科学院声学研究所基于概念层次网络理论的汉语理解系统。概念层次网络理论(Hierarchical Network of Concept,简称HNC)是中国科学院声学研究所研究员黄曾阳先生创立的面向整个自然语言理解的理论框架。这个理论框架是以语义表达为基础的,是以对语言的深层描写为出发点的,它对语义的表达是高度抽象化、

概念化、层次化、网络化的,所以称它为概念层次网络理论。人对语言的理解本质上是一种认知行为,如果能描述大脑认知结构的具体模式,计算机就可以运用这些模式对自然语言进行理解处理。HNC把认知结构分为局部和全局两类联想脉络,认为对联想脉络的表述是语言深层(即语言的语义层面)的根本问题。什么是局部联想和全局联想呢?简单地说,局部联想是指词汇层面的联想,全局联想是指语句及篇章层面的联想。更简单地说,理解句子有两种思路:一是从组成句子的词语入手,一是从句子的整体结构和上下文语境入手,前者就是局部联想,后者就是全局联想。当然,人在理解句子的时候,这两种联想不是截然分开的,而是并存的、相互作用的,计算机理解语言也应该综合运用这两类联想脉络。HNC的出发点就是通过建立两类联想脉络来“帮助”计算机理解自然语言。这种在语义表达上有自己的特色,在语义处理上走了一条新路的自然语言理解的理论框架HNC,抓住了汉字和汉语的特点,并绕开了现有汉语研究中的一些老大难问题,结果取得了可喜的研究成果,如不标调的拼音输入自动转换成汉字,以及分词连写拼音的研究均有较好的进展,1997年度还将推出语句理解的成果。

综上所述,汉语理解和汉语研究紧密相连。我国当前的汉语研究存在脱离汉语特点的问题。而我们必须看到,正确引导汉语理解和汉语研究,是事关我国人工智能开发的一个重要前题,难以回避。因此,我认为,汉语理解不应该建立在不符合汉语特点的、既有的汉语研究成果的基础上。以语义为基础、以对语言深层的描述为出发点的基于概念层次网络理论的汉语理解(HNC)系统可以避开当前汉语研究的困惑,符合汉语和汉字的特点,因而取得了汉语信息处理的可喜成果,其前景十分可观,值得引起汉语研究各方专家学者的注意。

主要参考文献

- [1]张志公:“语汇重要,语汇难”,中国语文,1988.1,另见《张志公文集》5卷
- [2]张志公:“汉语语法的再研究”,外语教学与研究,1990.3
- [3]刘宁:“张志公先生来校讲演”,北京外国语学院院刊,1990.4
- [4]黄曾阳:“HNC理论论文集”,中国科学院声学研究所声场声信息国家重点实验室自然语言理解课题组
- [5]张全:“基于HNC理论的语义块感知处理”,中国科学院声学研究所博士学位论文

(责任编辑 蔡德诚)