

自然语言处理中的一些宏观问题之我见

冯志伟 教育部语言文字应用研究所

摘要: 根据国家语委主任郝平在第三届“语言与国家”高层论坛暨第二届全国应用语言学系主任(所长)论坛上的主旨报告,本文提出加强自然语言处理技术和理论的研究,加强多语言信息处理技术的研究,加强跨学科语言信息处理人才的培养等建议。

主题词: 自然语言处理;多语言信息处理;跨学科人才

中图分类号: H08

文献标识码: A

文章编号: 1672-9382(2009)05-0036-07

2009年5月8日至11日,第三届“语言与国家”高层论坛暨第二届全国应用语言学系主任(所长)论坛在江苏省徐州市举行,教育部副部长、国家语委主任郝平出席论坛开幕式并做主旨报告。郝平指出,我国当前语言生活正快速发展变化,语言生活中各种矛盾凸显,社会需要提供语言服务的类型与方式与日俱增,汉语走向世界的脚步空前加快,争取国际话语权正成为民族的自觉意识。在此背景下,我国必须及时研究宏观语言战略,设计落实语言战略的行动计划,提出应对重大语言问题的科学预案。我们要从国家安全的角度作好语言规划,确定我国需要的关键语言,加大语言信息处理技术的投入,重点支持关键语言的信息技术,采取有效措施培养相关人才,保证国家和军事的信息安全。语言学家要对社会语言生活进行监测,并向社会公布语言生活状况;要跟踪研究语言生活中的热点问题,提出对策和建议。在谈到外语教育与外语的社会应用时,郝平指出,当今时代,文化成为民族凝聚力和创造力的重要源泉,成为综合国力竞争的重要因素。我们一如既往地做好“把世界介绍给中国”的同时,还要积极担负起“把中国介绍给世界”的新使命,架起中国与世界沟通交流的桥梁;中国应用语言学应加强与国外同行的合作;语言文字主管部门要研究外语生活,并加

强对外语应用的管理。

郝平的报告涉及应用语言学的各个方面,我个人的知识有限,只是对于应用语言学中的自然语言处理这一个领域还比较熟悉,这里,我从自然语言处理的角度,对自然语言处理中的一些宏观问题,谈谈我对郝平报告的理解。

1 加强自然语言处理技术和理论的研究

从传统的观点看,语言学和信息技术似乎是互不相关的两个不同的领域。语言学属于人文科学,信息技术属于技术科学,它们分别处于人类知识系统的不同方面,它们之间的界限似乎是泾渭分明的。然而,由于语言是信息的主要负荷者,计算机出现之后,由于信息处理技术的飞速发展,语言学和信息技术结下了不解之缘,出现了自然语言处理(Natural Language Processing,简称NLP)这个新的学科。

我们正处于一个信息网络的时代,在进入21世纪之后,几乎每一个生活在信息网络时代的现代人,都要直接或间接地与自然语言处理技术打交道。

在这个信息网络时代,科学技术的发展日新月异,新的信息、新的知识如雨后春笋般不断增加,出现了“信息爆炸”(information explosion)的局面。现在,世界上出版的科技

作者简介:冯志伟,教育部语言文字应用研究所研究员、博士生导师。研究方向:自然语言处理、计算语言学、机器翻译、应用语言学。E-mail: zwfengde@gmail.com。

刊物达 165 000 种, 平均每天有大约 20 000 篇科技论文发表。专家估计, 我们目前每天在因特网上传输的数据量之大, 已经超过了整个 19 世纪的全部数据的总和; 我们在新的 21 世纪所要处理的知识总量将要大大地超过我们在过去 2 500 年历史长河中所积累起来的全部知识总量。据 CNNIC 统计, 2002 年底全球的网页总数已经达到 10^9 这样的天文数字, 信息量的丰富大大地扩张了人们的视野, 人们希望能够准确地、迅速地获取到自己需要的信息, 自然语言信息处理技术, 已经成为了解决海量信息的获取问题的强有力的手段。

目前, 自然语言处理已经成为国际上最活跃的科学领域之一, 引起了全球科技界和企业界的密切关注。我国政府对于自然语言处理技术也非常重视, 投入了大量的经费。

在国家重大基础研究发展计划 973 项目中, 1999 年至 2003 年科技部首批立项的重大基础研究发展规划项目“图像、语音、自然语言理解与知识挖掘”将自然语言理解列为重要的研究内容, 在这个项目的支持下, 建立了中文语言数据联盟 (Chinese Language Data Consortium, 简称 Chinese LDC), 挂靠在中国中文信息学会, 其目标是建成具有完整性、规范性、权威性和系统性的通用中文语言资源库和中文信息处理评测体制, 为中文信息处理的基础研究和应用研究提供支持, 促进中文信息处理技术的发展。目前, 中文语言数据联盟有会员单位 70 多个, 各类语言资源 80 多种, 其中, 30% 的语言资源对会员免费提供, 在全世界范围内实现了中文语言数据资源的共享。该联盟自 2006 年正式运行以来, 每天都有专业人员进行网站访问和电话咨询, 已经共享语言资源 200 多套, 授权评测单位使用 40 多个, 在自然语言处理中发挥了很好的作用。2004 年科技部重大基础研究发展项目规划“数字内容理解的理论与方法”再次将自然语言处理作为重要内容, 其目的在于建立大规模的语料库、知识库和数据库, 作为语义计算 (semantic computation) 的基础, 在信息内容理解 (information content understanding) 的计算模型与方法方面, 研究信息内容理解的基础问题, 在给定需求的条件下进行语义计算; 在信息内容理解的关键技术和应用方面, 研究有害信息过滤和多媒体信息检索等国家有重大需求的基础应用技术, 建立计算模型和方法的验证环境。

国家“863”计划也投入了大量的资金用于自然语言处理技术的开发。2002 年的重大项目“奥运多语言智能信息服务系统关键技术及示

范系统研究”突出以人为本的信息服务, 通过网络手段向各国记者和观众提供综合、全面、多语种、可定制的信息服务, 使得任何人在任何时间和任何场合, 都可以获取奥运有关的信息, 从而通过“科技奥运”来实现“人文奥运”的目标。

国家自然科学基金委员会也支持自然语言处理的研究, 先后设立了重点项目、面上项目和青年基金项目, 研究范围涉及汉语、蒙古语、藏语、维吾尔语等语种语料库建设和语义分析等基础问题, 文字输入法、机器翻译、自动文摘等应用问题, 对自然语言的词汇、句子、语义、篇章等方面进行了有效的探索。1999 年的国家自然科学基金重点项目“汉语话语翻译关键技术研究”取得了具有创新性的重要成果, 建立了国际领先的多语种口语对照语料库, 研制了若干个有特色的实验口语翻译系统和多语种口语翻译平台。2007 年的国家自然科学基金重点项目“融合语言知识与统计模型的机器翻译方法研究”试图将基于规则的理性主义方法和基于统计的经验主义方法有效地结合起来, 提高机器翻译的质量。

国家哲学社会科学规划办公室也立项支持自然语言处理研究, 设立了相应的社会科学基金项目。2003 年立项的“计算语言学方法研究”, 总结了国内外的计算语言学方法, 使之系统化, 理论化, 具体化。由于方法的研究是自然语言处理系统 (诸如机器翻译、语料库、信息检索、信息抽取、文本分类等) 的关键问题, 这项研究成果对各种类型的自然语言处理实用系统的开发, 在方法上具有普遍的指导意义, 对于解决我国当前在自然语言信息处理中的理论和现实问题, 具有重要的推动作用。这个课题中总结出来的一些方法已经运用于中文信息处理的研究, 效果良好。

可以看出, 国家对于自然语言处理的大力支持, 促进了我国自然语言处理的发展。国家在我国自然语言处理技术的研制和发展中, 起了举足轻重的作用。

目前, 我国的自然语言处理已经取得了显著的成绩。语料库技术得到了充分的发展, 建立了一批具有重要影响的语言资源库, 面向信息处理的汉语基础研究有了长足的进展, 理论成果初见成效, 应用技术开发蓬勃发展, 产业化进程硕果累累。

我国开发的这些语言资源库和自然语言处理系统中, 部分技术已经达到或者基本达到实用化水平。例如, 各种类型的汉语语料库、现代汉语语法信息词典、知网、汉字输入系统、汉字激光排版系统、机器翻译系统、搜索引擎等。

许多新的研究方向不断出现，在实际应用的驱动下，自然语言处理技术不断与各种新技术相结合，开发出越来越多的实用技术。例如，网络内容管理和监控的研究，不仅与自然语言处理技术有关，而且与网络技术、情感计算、图像理解等技术有关；语音自动翻译技术涉及机器翻译、语音识别、语音合成、语音通讯等多种技术。

我认为，在我国的自然语言处理研究中，在基础资源的开发方面，目前还缺乏行之有效的国家标准和规范，至今我国还没有建立起权威性的国家语料库和语言知识库。在基础理论研究方面，还有许多重要的问题没有解决。例如，语义计算与形式分析问题、句法歧义问题、指代消解问题、未登录词的自动识别问题等。自然语言处理需要透彻地研制语言的形式模型，而我国在自然语言形式模型的研制方面还没有值得令人称道的成果。这些都是我国自然语言处理研究中亟待解决的问题，需要我們进行艰苦的探索。

这里，我想着重谈一下自然语言处理的形式模型研制的问题。

自然语言处理有着明确的应用目标，语音合成、语音识别、信息检索、信息抽取、机器翻译等，都是自然语言处理的重要应用领域。由于现实的自然语言极为复杂，不可能直接作为计算机的处理对象，为了使现实的自然语言成为可以由计算机直接处理的对象，在这众多的应用领域中，我们都需要根据处理的要求，把自然语言处理抽象为一个“问题”（problem），再把这个问题在语言学上加以“形式化”（formalism），建立语言的“形式模型”（formal model），使之能以一定的数学形式，严密而规整地表示出来，并且把这种严密而规整的数学形式表示为“算法”（algorithm），建立自然语言处理的“计算模型”（computational model），使之能够在计算机上实现。在自然语言处理中，算法取决于形式模型，形式模型是自然语言计算机处理的本质，而算法只不过是实现形式模型的手段而已。因此，这种建立语言形式模型的研究是非常重要的，它应当属于自然语言处理的基础理论研究。

由于自然语言处理的复杂性，这样的形式模型的研究往往是一个“强不适定问题”（strongly ill-posed problem），也就是说，在用形式模型建立算法来求解自然语言处理的问题时，往往难以满足问题求解的“存在性”、“唯一性”和“稳定性”的要求，有时是不能满足其中的一条，有时甚至三条都不能满足。因此，对于这样的强不适定性问题求解，应当加入适当的“约束条件”

（constraint conditions），使问题的一部分在一定的范围内变成“适定问题”（well-posed problem），从而顺利地求解这个问题。

自然语言处理是一个多边缘的交叉学科，因此，我们可以通过计算机科学、语言学、心理学、认知科学、人工智能等多学科的通力合作，把人类知识的威力与计算机的计算能力结合起来，给自然语言处理的形式模型提供大量的、丰富的“约束条件”，从而解决自然语言处理的各种困难问题。自然语言处理这个学科的边缘性、交叉性的特点，为解决这样的“强不适定问题”提供了有力的手段，我们有可能把自然语言处理形式模型的研究这个“强不适定问题”变成“适定问题”，这是我们在研究自然语言处理的形式模型的时候，值得特别庆幸的，也是应该特别注意的。

早在自然语言处理这个学科出现之前，语言计算研究的先驱者们就开始探索自然语言的形式模型。例如，Markov 链，Zipf 定律，Shannon 关于“熵”的研究，Bar-Hillel 的范畴语法，Harris 的语言串分析法，O.C.Кулагина 的语言集合论模型等。Markov 等具有远见卓识的学者很早就从形式描述的角度来研究自然语言，开自然语言处理形式模型研究的先河。

随着自然语言处理研究的发展，一系列的形式模型开始建立起来。这些形式模型大致可以归纳为如下几种：

（1）基于短语结构语法的形式模型：主要有 Chomsky 的短语结构语法，递归转移网络和扩充转移网络，自底向上分析法与自顶向下分析法，通用句法生成器和线图分析法，Earley 算法，左角分析法，CYK 算法，Tomita 算法，Chomsky 的管辖—约束理论与最简方案，Joshi 的树邻接语法等。

（2）基于合一运算的形式模型：主要有 Kaplan 的词汇功能语法，Kay 的功能合一语法，Gazdar 的广义短语结构语法，Shieber 的 PATR，Pollard 的中心语驱动的短语结构语法，Pereira 的定子句语法等。

（3）基于依存和配价的形式模型：主要有 Tesnière 的依存语法，德国学者的配价语法，Hudson 的词语法等。

（4）基于格语法的形式模型：主要有 Fillmore 的格语法和框架网络。

（5）基于词汇主义的形式模型：主要有 Gross 的词汇语法，Sleator 和 Temperley 的链语法，词汇语义学，词网（WordNet）等。

（6）基于概率和统计的形式模型：主要有 N-元语法，隐马尔可夫模型（Hidden Markov Model，简称 HMM），最大熵模型，条件随

机场 (Condition Random Field, 简称CRF), Charniak 的概率上下文无关语法和词汇化的概率上下文无关语法, Bayes 公式, 动态规划算法, 噪声信道模型, 最小编辑距离算法, 决策树模型, 加权自动机, Viterbi 算法, 向前算法等。

(7) 语义自动处理的形式模型: 主要有义素分析法、语义场理论, 语义网络理论, Montague 的蒙塔鸠语法, Wilks 的优选语义学, Schank 的概念依存理论, Mel'chuk 的意义—文本理论等。

(8) 语用自动处理的形式模型: 主要有 Mann 和 Thompson 的修辞结构理论, 文本连贯中的常识推理技术等。

可以看出, 我国学者在上面列举的这一系列的形式模型中, 基本上还没有什么突出的贡献。因此, 我认为, 在注意自然语言处理的应用研究的同时, 亟待加强自然语言处理的形式模型的研究, 为世界的自然语言处理形式模型的研究, 做出我们中国学者应有的贡献。

这些自然语言处理的形式模型, 都从自然语言的某个侧面揭示了它的某些形式特性, 但是还不能从根本上揭示自然语言的全部形式特性。自然语言处理是一个非常复杂的问题, 迄今为止, 我们对于自然语言的形式特性了解还远远不够, 还有许多的问题需要我们去探索。我研究自然语言处理已经 50 多年了, 我今年已经是年过古稀的老人, 可是, 自然语言处理还是一个非常年轻的学科, 她还是一棵正在成长中的枝叶茂密的大树。我们个人的生命是有限的, 而自然语言处理这棵科学的大树却是永远长青的。我们个人的生命与自然语言处理这棵科学的大树相比, 显得多么渺小! 庄子说, “吾生有涯而知无涯, 以有涯求无涯, 不亦殆乎?” 我个人认为, 庄子的这种说法是不对的。在自然语言处理的研究中, 我们是以有涯的生命来探讨无涯的知识, 充满了无比的乐趣。我们应该把庄子的话中的“殆”字改为“乐”字, 把这句名言改为: “吾生有涯而知无涯, 以有涯求无涯, 不亦乐乎?” 让我们在探索自然语言处理奥秘的研究工作中, 做出我们自己的贡献, 从中获取最大的乐趣, 从而体现我们生命的价值。

总而言之, 加强自然语言处理的研究, 是我国应用语言学研究特别值得关注的一个大问题。郝平在报告中强调, 要“加大语言信息处理技术的投入, 重点支持关键语言的信息技术”, 这是有远见卓识的。

2 加强多语言信息处理的研究

随着对外交流的发展, 不同语言之间的信

息交换显得越来越重要, 互联网实际上已经成为一个多语言的网路。在这个多语言网络的年代, 中国语言文字的信息化必须解决多语言的信息处理 (multilingual processing) 问题。

语言是信息的最主要的负荷者, 如何有效地使用现代化手段来突破人们之间的语言障碍, 成为全人类面临的共同问题。多语言信息处理技术包括机器翻译技术、跨语言信息检索技术、多语言问答式信息检索技术、多语言信息资源建设技术, 这些技术是解决语言障碍问题的有力手段。由于自然语言是极端复杂的, 多语言信息处理技术又涉及多种语言, 因此就更加复杂和困难。

语言对于人类交流思想的重要作用是毋庸置疑的。为了克服语言的障碍, 曾经有人提出使用人类通用语言 (lingua franca) 来替代各种不同语言的想法。但是, 这样的想法显然是很难实现的, 就是有了这样的通用语言, 它也代替不了各种不同的民族语言。因为语言是民族文化的象征, 放弃民族语言就意味着放弃民族的文化, 如果全人类都讲一种通用的语言, 各具特色的、丰富多彩的民族文化也就黯然失色了。这显然不是一件好事。尽管英语在英国和美国占绝对的统治地位, 在英国的威尔士, 人们还在讲威尔士语, 在美国的某些地区, 还有很多人在讲西班牙语。至于像加拿大和瑞士这样的双语和多语国家, 像欧洲联盟和联合国这样的组织, 多语言的使用 (multilingualism) 已经成为日常生活中的基本原则和普遍现象。欧盟委员会在 2005 年 11 月 22 日公布了一个题为“实现多语系策略”的官方报告, 这份报告的题记使用了斯洛伐克的一句谚语: “你懂得的语言越多, 你就越像一个人”。这句谚语成为该报告的基调。可见多语言的使用已经成为欧盟的一个众人瞩目的大问题。而在多语言的使用中, 多语言信息处理技术的需求会变得越来越迫切和尖锐。

我们认为, 面对多语言问题, 可以使用社会的手段来解决, 也可以使用技术的手段来解决。我这里只谈谈如何使用技术手段来解决多语言问题。

使用技术手段, 可行的解决策略如下:

(1) 通过机器翻译来解决多语言问题

机器翻译 (machine translation) 要使用电子计算机把一种语言 (源语言, source language) 翻译成另外一种语言 (目标语言, target language)。它涉及语言学、计算机科学、数学等众多学科, 是非常典型的多边缘的交叉学科。

机器翻译研究一直是一个全球性的课题, 欧美各国和日本对机器翻译研究的兴趣在持续增长, 比较著名的翻译系统有加拿大的 METEO

系统, 欧共体的 Eurotra 系统, 日本政府资助的 ODA 系统, 日本 NEC 公司的 PIVOT 系统, 日本富士通公司的 ATLAS 系统, 荷兰的 Rosetta 系统和 DLT 系统, 以及美国的 Systran 系统以及 Carnegie-Mellon 大学的 KBMT 系统等。

我国从 50 年代开始研究机器翻译问题, 当时只有几个研究所和大学参加, 仅限于俄汉翻译。在国家的支持下, 经过 40 多年, 特别是近 20 年的努力, 据不完全统计, 已有四、五十个单位从事机器翻译的理论和实验研究, 也与多个国家和地区开展了各种形式的合作。我国机器翻译研究取得了多项成果, 翻译软件的产品也多达十几种, 涉及英汉、汉英、日汉、汉日、俄汉、德汉、法汉、英日等, 其中比较著名的翻译产品有华建集团的英汉机译系统、中软总公司的“译星”英汉翻译系统等。这些都标志着我国翻译软件已有了极大的发展。但是, 我国机器翻译的正确率还不高, 不论是译文对原文的忠实度 (fidelity), 还是译文本身的可懂度 (comprehensibility), 离真正实用还有相当大的距离。

机器翻译主要可以分为基于规则的机器翻译 (rule-based machine translation) 和基于语料库的机器翻译 (corpus-based machine translation) 两种, 基于语料库的机器翻译又可以进一步分为基于统计的机器翻译 (statistics-based machine translation) 和基于实例的机器翻译 (example-based machine translation)。为了提高机器翻译的正确率, 我们应当在消化前人已有研究成果的基础上, 把基于规则的机器翻译和基于语料库的机器翻译巧妙地结合起来, 建立多语言语料库, 对多语言语料库在句法、语义的层次上进行深加工, 进一步建立多语言树库 (multilingual tree bank), 使用机器学习 (machine learning) 的技术从多语言树库中获取统计性的翻译规则, 再利用这样的统计性翻译规则来进行机器翻译。基于规则的方法和基于统计的方法的巧妙结合, 将使我国的机器翻译提高到一个新的水平。

(2) 通过跨语言信息检索来解决多语言问题

传统的信息检索系统 (Information Retrieval, 简称 IR) 主要是在一种语言之内进行的, 用户使用自己最熟悉的语言作为查询语言, 来检索用这种语言书写的文本。随着网络上不同语种的增加, 用户面对查询多语言信息的情形越来越普遍, 这就出现了以一种语言描述用户的查询意图与网络上不同语言的文本相互匹配的问题, 也就是需要跨过不同的语言来进行检索, 这就是“跨语言

信息检索” (Cross-Language Information Retrieval, 简称 CLIR)。跨语言信息检索的问题不仅在互联网上存在, 在跨国公司、国际组织 (如联合国、欧盟)、大型的国际活动 (如奥运会、世博会) 中, 也存在跨语言信息检索的迫切需求。

跨语言信息检索的技术难点是索引处理、匹配策略、排序策略、搜索策略、反馈机制和查询扩展技术, 我们应当研究这些技术, 提高跨语言信息检索的易用性、查全率和查准率。

(3) 通过多语言问答式信息检索来解决多语言问题

目前的信息检索是按照关键词来进行查询的, 用户根据自己的查询意图, 使用关键词或者布尔表达式提问, 系统根据相关性大小的顺序, 返回与用户提问相关的网页链接, 用户逐一访问这些链接, 最后找到自己最满意的查询结果。“多语言问答式信息检索” (multilingual question answer information retrieval) 与关键词查询不同, 它允许用户以自己熟悉的自然语言问句向系统提问, 而不使用关键词或布尔表达式, 这样不仅可以精确地表达查询意图, 而且符合人们的习惯, 使得信息检索向人性化、智能化的方向发展。多语言问答式信息检索要在多语言的文档中进行搜索, 可以直接返回答案或者蕴涵答案的文本片断, 免除了人们通过链接继续查询的麻烦, 从而提高了信息检索的效率。

文本检索会议 (Text Retrieval Conference, 简称 TREC) 是由美国国家标准局 (National Institute of Standard and Technology, 简称 NIST) 组织的一年一度的信息检索国际会议, 从 1992 年开始已经召开了 13 届, 其目的在于推动大规模的文本检索研究, TREC 提供一个标准的文档库以及一套评测方法, 为文本检索建立了一个公平竞争的平台。在 TREC 的专题 (Track) 中就有一个“问题回答专题” (Question-Answer Track), 可见多语言问答式信息检索已经成为国际学术界关注的焦点。

多语言问答式信息检索的技术难点是自然语言问句的句法和语义结构的形式表示方法以及自动分析技术, 我国在这些技术上有多年的积累和丰富的经验, 我们要进一步研究这些技术, 以便准确地表达用户使用自然语言提问的内容, 提高多语言问答式信息检索的易用性和效率。

多语言问答式信息检索系统还应当支持各种常用的文件类型, 如 HTML 文件、图像文件 (JPEG, GIF)、ASCII 文件、URL 文件、Word 文件、Excel 文件、Powerpoint 文件、PDF 文件等。

(4) 通过多语言信息资源建设来解决多语言问题

机器可读词典和语料库是最重要的语言资源,它们是语言信息处理的粮食,没有这些资源,语言信息处理就成了无米之炊。“多语言信息资源建设”(multilingual information resources construction)的主要目标是:

建立机器可读的多语言对译词典,我们应当对现有的各种机器可读词典进行集成,使之成为多语言信息处理的宝贵资源。

建立英语—汉语双语语料库,开展双语对齐技术的研究。目前,北京大学计算语言学研究所的英语—汉语双语语料库中,对齐的句子已经有5万对,并且开发了相应的对齐工具和管理软件。我们应当在这样的基础上进一步扩充和完善这个双语语料库。

建立多语言并行语料库,使之成为多语言机器翻译的重要资源。

郝平在报告中指出,当今时代,文化成为民族凝聚力和创造力的重要源泉,成为综合国力竞争的重要因素。我们一如既往地做好“把世界介绍给中国”的同时,还要积极担负起“把中国介绍给世界”的新使命,架起中国与世界沟通交流的桥梁。因此,加强外语应用和多语言信息处理的研究,是一件具有战略意义的大事。

3 加强跨学科语言信息处理人才的培养

一般地说,计算机对自然语言的研究和处理应当经过如下4个过程:第一,从语言学的角度提出自然语言处理的问题和理论;第二,把需要研究的语言学问题加以形式化,使之能以一定的数学形式或者接近于数学的形式,严格而规整地表示出来;第三,把这种严格而规整的数学形式表示为算法,使之在计算上形式化;第四,根据算法编写计算机程序,使之在计算机上加以实现。

因此,为了从事自然语言处理的研究人员不仅要具备语言学的知识,而且还要具备数学和计算机科学方面的知识。这样,自然语言处理就成为一门介于语言学、数学和计算机科学之间的边缘性交叉学科,它同时涉及文科、理科和工科三大领域,这使它具有明显的跨学科性质。

我们希望自然语言处理的研究人员同时具备语言学、数学和计算机科学的知识,成为文理兼通、博学多识的人才,不过,这种人才的培养是非常不容易的。我觉得,在缺少文理兼通的跨学科人才的情况下,我们可以退一步这样来要求:对于不可能同时具备语言学、数学和计算机科学知识的研究人员,至少对于自己原来所学的专业

是精研通达的内行,对于另外两个专业不是似懂非懂的外行;在研究工作中,不同背景的人员互相学习,取长补短,这样,也有可能有效地从事自然语言处理的研究工作。

我们应该提倡自然语言处理的研究人员不断地进行更新知识的再学习。“活到老,学到老”,对于自然语言处理研究人员来说,决不是一句装扮门面的空话,而应该成为身体力行的座右铭。

由于我国中学教育实行文理分科的办法,我国文科背景的研究人员数学和计算机知识都显得不足。为了培养跨学科的自然语言处理人才,我们需要进行教材建设,出版一些帮助文科背景的人员提高他们的数学和计算机科学水平的书籍。

郝平在报告中指出,我们在重点支持关键语言的信息技术的同时,还要“采取有效措施培养相关人才”。这是具有长远战略意义的真知灼见。□

参考文献

- [1] Barbara H. Partee. 冯志伟导读. *Mathematical Methods in Linguistics* [M]. 北京:世界图书出版公司,2009.
- [2] D. Jurafsky & J. Martin. 冯志伟,孙乐译. 自然语言处理综论 [M]. 北京:电子工业出版社,2005.
- [3] 冯志伟. 机器翻译研究 [M]. 北京:中国对外翻译出版公司,2004.
- [4] 冯志伟. 应用语言学新论——语言应用研究的三大支柱 [M]. 北京:当代世界出版社,2003.
- [5] 冯志伟. 自然语言处理的形式模型 [M]. 合肥:中国科学技术大学出版社,2009.
- [6] 刘海涛. 欧洲联盟语言状况和语言政策 [A]. 中国语言生活状况报告(2005) [C]. 北京:商务印书馆,2006.
- [7] 宗成庆,高庆狮. 中国语言技术进展 [J]. 中国计算机学会通讯. Vol. 4, No. 8, 2008: 39-48.

Some Macroscopic Opinions on Natural Language Processing

Abstract: Based on topic speech of Hao Ping (the Chairman of State Language Commission) in Third Summit Forum on “Language and Country” and National Forum of Directors of Applied Linguistics Institute, the author proposed his opinions on the research of technology and theory in natural language processing, on the research of multilingual information processing, and on the training of cross-discipline personnel.

Key Words: natural language processing; multilingual information processing; cross-discipline personnel