《现代语文》2006 年第 12 期（总第 258 期）

关于信息时代的多语言问题的一些思考

冯志伟
教育部语言文字应用研究所

摘要：我们正处于一个多语言网络时代，如何有效地使用现代化手段来突破人们之间的语言障碍，成为了全人类面临的共同问题，这就是多语言问题。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是解决多语言问题的技术手段，多语言移动电话热线服务则是解决多语言问题的社会手段。

关键词：多语言问题，机器翻译，信息检索，问答系统，信息资源，移动电话服务。

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络，我们正处于一个多语言网络的时代，因此，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。

现在我们已经进入了信息网络的时代，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，成为了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，这些技术是解决语言障碍问题的有力手段之一。由于自然语言是极端复杂的，多语言信息处理技术又涉及到多种语言，因此就更加复杂和困难。

可以毫不夸张地说，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多语言信息处理技术打交道。不论对于社会政治还是对于经济发展，多语言信息处理技术都无疑是一个重要的研究领域。

从社会政治上来说，语言对于人类交流思想的重要作用是毋庸质疑的。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）已经成为了日常生活中的基本原则和普遍现象。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。

我们以欧洲联盟（以下简称欧盟）的多种语言翻译为例来说明这个问题。欧盟作为一个共同体，成功地建立了统一大市场，促进了各成员国的经济大发展；启动了欧洲统一货币，形成了横跨欧洲大陆的“泛欧元区”，足可以同美元抗衡；基本实现了共同边界，建立了共同外交和安全政策；现在正在讨论制定欧洲宪法，进一步加速欧洲经济、政治一体化的进程。但是语言问题一直困扰着这个国际组织，而且至今还没有应对的良策。按照欧盟的前身欧洲经济共同体 1957 年首脑会议通过的关于语言多样化的决议，各成员国都有平等使用本国语言的权利，各成员国的官方语言即成为这个组织的工作语言。当时只有 6 个创始国，开会时只使用 4 种语言，语言问题还不十分突出。1995 年欧盟扩大到 15 个国家，官方语言增加到 11 种（比利时、奥地利没有自己的独特的语言，爱尔兰官方语言为英语、爱尔兰语，卢森堡因其官方语言为德语、法语和卢森堡语，没有要求使用卢森堡语），语言问题开始突显出

来。按照规定，在召开欧盟的“三驾马车”欧洲议会、欧盟首脑会议、欧盟委员会等正式会议时，这 11 种语言就是会议的工作语言。也就是说，欧盟在召开正式大会时应当提供 11 种语言的翻译服务，大会文件要译成 11 种语言，由此增加的翻译工作量之大，财政支出之高是可以想见的。据统计，欧盟每个正式会晤至少需要 33 个翻译人员，才能完成 11 种语言间高达 110 种可能的互译。据媒体报道，每年欧盟大小会议共 1 万多个，为此要付出 15 万个翻译人/日，全年翻译的文件达 113 万页以上。每次为期 5 天的欧洲议会起码需要 450 名翻译人员。目前欧盟雇用的常年翻译有 460 人，临时翻译 1 500 多人。即便如此，仍不能满足所有成员国的所有语言要求，这支庞大的翻译队伍如今越来越难以应付巨大的工作压力，已陷入捉襟见肘和穷于应付的困境。据报道，2003 年欧盟的文件翻译总量有 148 万页之多。不用说，翻译方面的支出也是巨大的，据媒体披露，整个欧盟目前一年用在翻译方面的支出高达 5.5 亿欧元，每年用于翻译的支出约占全部行政管理预算的一半左右。语言的困扰随着欧盟成员国的增加而愈加尖锐：一方面，所有成员国的代表当然有权使用本国的语言表达各自的意见，另一方面，由此增加的翻译工作量和财政支出却并非是所有成员国都乐于承担的。从 2004 年 5 月 1 日开始，欧盟成员国已经扩大到 25 个，这就意味着欧盟正式会议的工作语言至少有 20 种，一场会议的同声传译至少需要 60 名翻译人员。欧盟每天都有七八百名翻译人员在进行同声传译工作。现在欧盟就因同声传译跟不上，笔译错误层出不穷而叫苦不迭，一旦 20 种官方语言同时互译，其后果将更加混乱，而用于翻译方面的财政支出将大幅度提高，据报道，2004 年由于 10 个新成员国的加入，文件翻译总量已经达到 206.5 万页，2005 年更是将高达 237 万页，其支出将增至 8.08 亿欧元，对欧盟各成员国无疑会增加更大的负担，在语言方面带来更多的矛盾与困扰。以至于在 2004 年 5 月，欧盟不得不通过一项新的议案，规定所有欧盟成员国在欧盟会议上的文件都不得超过 15 页 A4 纸，以减轻翻译人员的工作量。在我们这个多语言的世界上，翻译的重要性是不言而喻的，而随着人类交际活动的发展，翻译的数量越来越多，翻译的需求将越来越迫切，人工的翻译已经远远满足不了社会的需求，在这种情况下，作为多语言信息处理技术的一个分支的机器翻译必然会成为人工翻译的一个重要的补充而发展起来。

对于经济发展来说，在这个多语言的世界上，翻译对于推销商品的重要性是众人皆知的。如果中国的工业产品在美国市场上销售，美国人当然希望这种产品的说明书是用英文写的，而不是用中文写的。如果德国的药品在中国市场上销售，中国的顾客当然也希望这种药品的说明书是用中文写的，而不是用德文写的。翻译是一种高智能的劳动，它要求熟练的翻译技巧，它需要丰富的语言知识和专业知识，因此，翻译的开销是很高的。如果采用多语言信息处理技术中的机器翻译来减轻翻译人员的负担，提高翻译工作的效益，它的经济价值也是很高的。根据国际权威机构对于世界翻译市场的调查显示，翻译市场的规模在 2000 年已经上升到 130 亿美元，在 2005 年将达到 227 亿美元，而中国的翻译市场将达到 200 亿人民币。随着互联网应用范围的扩大和国际电子商务市场的日渐成熟，到 2007 年，只是网页的翻译业务将达到 17 亿美元的规模。目前，我国翻译能力严重不足，我国翻译市场的规模尽管已经超过了 100 亿人民币，但是现有的国内翻译公司只能消化 10%左右，由于无法消化大量从国际上传来的信息流，我们的信息不灵，就有可能使我们在国际竞争中失去大量的商业机会。传统的人工翻译已经难以满足实际的需要，在这种情况下，多语言机器翻译技术将可能改变我国这种翻译能力严重不足的局面。另外，跨语言信息检索技术、多语言问答式信息检索技术以及多语言信息资源建设对于经济发展也是非常重要的。

随着我国申请 2008 年在北京举办奥运会以及 2010 年在上海举办世博会的成功，届时操不同语言的各国运动员、政府首脑、著名人士、企业家、新闻记者以及世界各地的数以万计的旅游者将来我国参加这样的大型国际活动，多语言问题必将显得特别突出；如何解决多语言问题，使来自世界各地的人们能够使用各自的母语轻松地进行交流，是我们应当认真地考

虑的问题。

我们认为，面对多语言问题，可以使用技术的手段来解决，也可以使用社会的手段来解决。

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）。它涉及到语言学、计算机科学、数学等许多部门，是非常典型的多边缘的交叉学科。在语言学中，机器翻译是计算语言学的一个研究领域；在计算机科学中，机器翻译是人工智能的一个研究领域，在数学中，机器翻译是数理逻辑和形式化方法的一个研究领域。机器翻译要把这些不同的学科结合起来，综合地进行研究。机器翻译要求不同学科的专家通力合作，取长补短，相得益彰。

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 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，经过四十多年，特别是近二十年的努力，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，涉及的语种有英汉、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中软总公司的“译星”英汉翻译系统、华建集团的英汉机译系统、北京邮电大学智能科学技术研究中心的“面向奥运的多语种人机对话智能平台与智能移动终端系统”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有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，对于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，用户使用他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）是由美国国家技术标准局（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、图象文件（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（汉语、英语、日语、韩国语、藏语、蒙语、维吾尔语），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。

前面 4 种解决策略都是使用技术手段，而多语言移动电话热线服务则是解决多语言问题的社会手段。

在 2008 年举办奥运会或者在 2010 年举办世博会的时候，我们建议奥运会或世博会的领导者可以征集翻译志愿者组成强大的多语言翻译服务队，参加多语言翻译服务队的志愿者至

少应该精通一门外国语，根据他们所懂外语的不同组成不同语种的志愿翻译小组，翻译志愿者提供出他们各自的手机号码，每一个语种的志愿翻译小组提供出该语种的每一个志愿翻译者的手机号码，把这些手机号码印刷在一个翻译热线问询卡上，这样，我们就可以有英语的翻译问询卡、日语的翻译问询卡、德语的翻译问询卡、法语的翻译问询卡，…，等等。当操不同语言的外宾来到中国的时候，根据他们语种的不同发给他们不同的翻译问询卡。这样，当外宾遇到多语言交际困难的时候，就可以拨打相应的手机号码，向翻译志愿者求助，翻译志愿者向他们提供免费的翻译问询服务。如果在外宾拨打手机时对方无应答或者已经关机，可以拨打其他翻译志愿者的手机号码。只要在奥运会或世博会期间，我们组织翻译志愿者提供连续的不间断的服务，一定能满足外宾的要求。这样的多语言移动电话热线服务，或许是解决奥运会或世博会多语言问题的可行手段。2002 年我到韩国访问，正值世界杯足球赛（FIFA），韩国就使用了翻译热线电话进行翻译问询服务，效果很好。下面是我在仁川机场得到的一张汉语的翻译问询卡（这是正面，背面是一些提供翻译问询服务的手机号码）：

使用这样的翻译问询卡，有效地解决了语言障碍问题。

当然，这只是我个人的意见，这个意见没有政府和领导的支持，是不可能实现的，希望有关领导部门能够注意我们的这个意见。在我国的多语言技术还不十分成熟的时候，这样的办法实行起来并不很难。

参考文献

1. 冯志伟，机器翻译研究[M]，中国对外翻译出版公司，2004 年。
2. 冯志伟，应用语言学新论 — 语言应用研究的三大支柱[M]，当代世界出版社，2003 年。
3. D. Jurafsky，J. Martin，自然语言处理综论（冯志伟、孙乐译）[M]，电子工业出版社，2005 年。