民文机器翻译所面临的瓶颈问题-资源与技术

那顺乌日图

一、简要的回顾

民文机器翻译研究，从严格意义上来讲，始于上世纪九十年代末。虽然从80年代开始在国内外曾有人提出或试图做民文机器翻译研究，但当时由于理论、技术上的条件都不成熟，从而一直没有得到实施。到九十年代后期，民文信息处理工作经过近二十年的探索，基本上走过了字处理阶段。这样，人们在搞基础理论研究的同时开展应用系统的开发。其中机器翻译大概是除了电子出版以外的另一个热点。

- 青海师范大学是进行汉藏机器翻译较早的单位，他们在国家863计划资助下开发汉藏翻译系统及电子词典，引起学界和用户的广泛关注，其开发的汉藏（藏汉）在线翻译和智能化汉藏（藏汉）翻译系统都得到广泛应用；

1998年内蒙古大学蒙古学学院、中国科学院计算技术研究所、北京大学计算语言学研究所共同承担国家863项目“面向政府文献的汉蒙机器辅助翻译系统”，研制出“达日罕系统-政府文献版V1.0”，此后上述三个单位又承担一期863项目对该系统进行改进；

2009年新疆大学与新疆信息产业有限公司合作在工信部电子信息产业发展项目的资助下开展了汉-维、柯辅助翻译软件的研发；

2010年中科院计算所与新疆大学合作推出了基于统计的维汉机器翻译系统；

2011年中科院新疆理化所也推出了汉维/维汉统计机器翻译原型系统；

据不完全统计，目前进行民文机器翻译的除了上述几个单位之外至少还有：

中科院计算所（汉-蒙、藏、维，蒙、藏、维-汉）

中科院智能所（汉-蒙、？）

西北民族大学（藏-汉、？）

中科院自动化所（藏、(新/老)维、蒙-汉）

清华大学（蒙、藏、维-汉）

哈尔滨工业大学（汉-维、？）

北京理工大学（？）

内蒙古师范大学（蒙-汉）

…… ？

2009年CWMT首次将汉蒙机器翻译列入测试范围，7个单位的翻译系统参加测试，2011年CWMT又将测试语言范围扩展到藏、维等语言和蒙汉机器翻译。

建立面向少数民族地区和周边国家的双语或多语网站，开发少数民族语言文字多媒体教学系统等，都将需要民族语言机器翻译系统。

但是目前大多系统的水平，尚未达到能够很好地满足用户要求的、产品化的程度。

- 蒙古国相关大学和研究机构也在开发英蒙，蒙英……等。
- 日本諏訪東京理科大学 等也在开发日-蒙机器翻译系统。
- 在国外开发藏文、维文系统的也不乏其人……

- 那么民文机器翻译系统发展缓慢的主要瓶颈在哪里？
- 这个问题还得与机器翻译技术本身的发展结合起来考虑：

90年代基于规则的机器翻译技术比较盛行的时候，民文翻译的效果还算可以，因为当时基于规则的系统准确率基本上都不高；

- 2000年后基于实例的机器翻译开始盛行的时候，民文机器翻译的发展速度就受到了制约，那是因为没有一个令人满意的，哪怕是满足基本需求的实例库都很少；
- 近几年基于统计的机器翻译作为机器翻译主流技术的时候民文机器翻译同样发展缓慢，主要瓶颈还是资源匮乏 。

如何解决资源问题？

建立资源联盟？

申请专项？

还请大家多加关注。

谢谢！