在国际互联网中消除中文同外文之间语言壁垒的研究

[1,2]董名垂　[1,2]黄辉　[2]费晓磊　[2]窦佳易

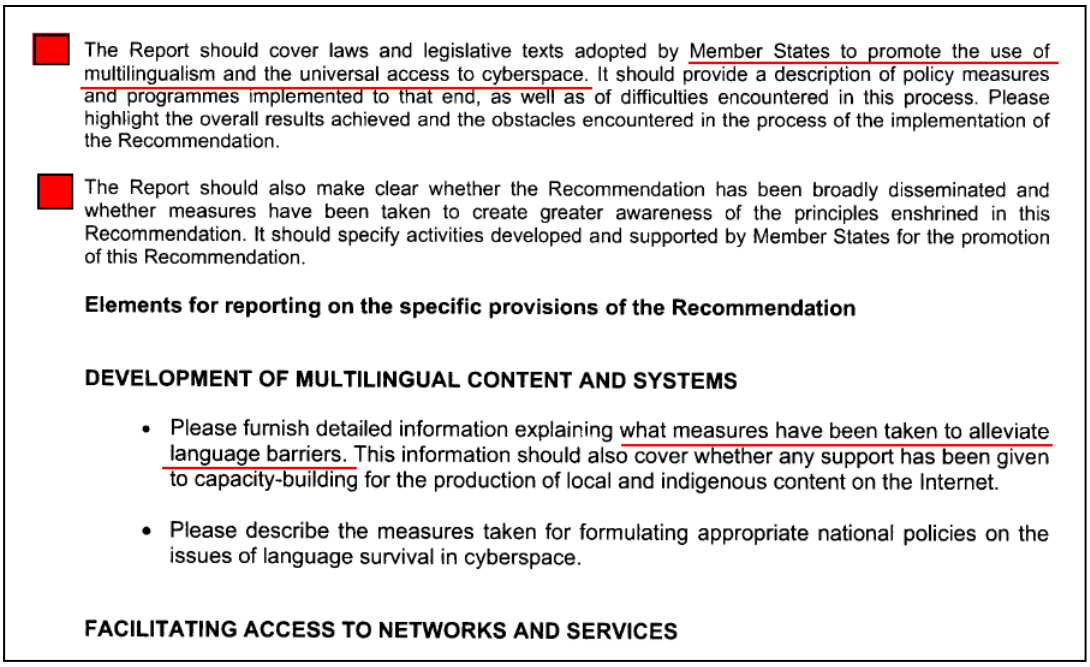
[1]澳門电脑与系统工程研究所　[2]澳門大学

{dmc, derek}@inesc-macau.org.mo; {ma66558, da72815}@umac.mo

摘要

联合国科教文组织 UNESCO 要求其所有成員国自 2006 年開始每四年向 UNESCO 汇报一次各成员国在 Internet 中减低多语言壁垒，加强民众存取国际互联网资料所采取的有效措施及取得的成效。图 1 示出了 UNESCO 下发官方文件中的有关内容。澳门政府委托澳门大学作出调研并于 2006 年 12 月写就这份提交到联合国科教文组织的大型报告。在报告中汇报了澳门政府为实施上述要求业已制订的法律法规以及产生的效果，详细列出了澳门为加强民众存取国际互联网信息，有效推广电子政务、电子商务、电子事务三方面所投放的巨额研发资金，实施的具体措施以及目前的现状等统计资料。特别强调指出，部分澳门高校及研究机构远见卓识，为解决在 Internet 虚拟世界内多语言壁垒问题，於十多年前就已开展了解决中文同外文之间电子化互译技术的研究，并已研究开发出了初步的样机系统。基于这一成果，在报告中建议 UNESCO 这一国际组织领头召集协调各有兴趣的成员国一起合作研究解决国际互联网多语言壁垒问题。希望经过数年努力，在 Internet 虚拟世界内实现多语言无疆界、无阻隔，为各国民众带来无障碍读懂国际互联网资料的地地道道的实惠，更为弘扬中华文化，为实现中文国际化、全球化作出贡献。本文前两位作者参与并主持了该份大型报告的书写。基于该报告的思路，本文将介绍有关中文同外文之间在线机助翻译、网页在线即时翻译、在线网上实时互译互通技术，並展示初步的实测結果。目的只有一个，就是希望技术、愿望、实践能整合为一体，早日实现中文同其他外文在线网上实时互译互通。

关键词：单机版机助翻译，在线机助翻译，网页在线即时翻译，在线网上实时互译互通



The Report should cover laws and legislative texts adopted by Member States to promote the use of multilingualism and the universal access to cyberspace. It should provide a description of policy measures and programmes implemented to that end, as well as of difficulties encountered in this process. Please highlight the overall results achieved and the obstacles encountered in the process of the implementation of the Recommendation.

The Report should also make clear whether the Recommendation has been broadly disseminated and whether measures have been taken to create greater awareness of the principles enshrined in this Recommendation. It should specify activities developed and supported by Member States for the promotion of this Recommendation.

Elements for reporting on the specific provisions of the Recommendation

DEVELOPMENT OF MULTILINGUAL CONTENT AND SYSTEMS

- Please furnish detailed information explaining what measures have been taken to alleviate language barriers. This information should also cover whether any support has been given to capacity-building for the production of local and indigenous content on the Internet.
- Please describe the measures taken for formulating appropriate national policies on the issues of language survival in cyberspace.

FACILITATING ACCESS TO NETWORKS AND SERVICES

圖 1. UNESCO 下发官方文件的有关内容

1. 解决网页多语种文字互通的传统方案

灿烂的中华文化其内容博大精深、其历史渊源流长，但在Internet国际互联网世界却不得不面对比别国文字更多的问题，即对内要整合沟通因历史原因造成的两岸四地繁简体、内码及用词不同的问题，对外要应对中文切词断句同外文互译更为困难的技术难题。特别是后者，随着国际互联网的普及使用，令更多民众消除地域、时间阻隔，融入到Internet虚拟世界，可以不费吹灰之力就查询到他们想要获得的世界各国各地多媒体信息，然而现阶段最大的阻隔依然是语言的壁垒，使得民众从Google或Yahoo这样一些国际知名信息网站上轻易查到了所要的信息却看不懂用别国文字书写的内容；同样地，不懂中文的外国人也在为看不懂中文网站的丰富内容而苦恼。

解决这一难题的传统做法就是在网站上开设多语言选择功能，其实质就是把同样的网页内容事先复制并翻译到其他语种，编排好了存放在后台网络服务器数据库内，当前台用户查询该网页并选定了语种时，电脑就按用户选定的语种调出该语种版本供用户阅读观看。如此做法的最大弊端有两个，一是同样的海量信息内容必须按照使用的语种数量被多次重复复制，不仅成倍增加人们翻译

制作网页信息的工作量，更可怕的是导致信息量成倍冗余爆炸，大量浪费有限的存储资源；二是很多网页信息是在不断动态变换的，复制并人工翻译、排版后再输入到电脑显然丢失了网页的动态实时性，采用上述模式在实践和理论上都无法实现动态实时替换网上的变化信息。基于此，快速的机助自动翻译技术在七十年代末就引发了人们的强烈兴趣并因此进入了漫长的研究开发岁月。

2. **在线机助翻译**

为了提高机助翻译的精度和速度，人们先后研究提出了基于词汇语义分析的、基于翻译规则的、基于词库案例的、基于翻译统计规律的、基于上下文语义/词法/句法/语法分析的、基于形态和语义特征分析的等等各种机器自动翻译技术，并加入了机器增量式学习功能和各种人工智能。随着机助翻译技术的不断进步，从九十年代初起，一批商品化的软硬件电子字/词/辞典；逐词、逐句、逐段、整段机助翻译软件被纷纷推向了市场。在两岸四地都有不少高校和研究机构，例如北京清华大学、北京大学、中国科学院、香港中文大学、澳门电脑与系统工程研究所、台湾新竹清华大学、台湾欧泰科技 OTEK 公司等单位在与时俱进地研究开发中文同其他外文之间的机助翻译技术及系统，双向互译的语种已经有十二、三种之多，经过长达二十多年的研究开发，一些单机版机助翻译软件系统在市场上售出并获得应用。但限于电子化中文是世界上最难处理的语言之一，所以两岸四地对机助翻译技术的研究多半还滞留在单机版本逐词查询翻译，以及逐句或整段文字的机器自动翻译上，到目前为止，尚未看到有出自两岸四地的在线机助翻译产品在国际互联网上使用。

最近几年，应映广大网民的需要，Google、Yahoo、金桥在线等网站继推出网上电子字/词/辞典后，又相继推出了在线机助翻译功能。甚至在微软公司零三年版本的 Word 软件中就已嵌入了 WorldLingo 公司的系统，能实现在线人机交互式整段翻译，供用户选用的可双向翻译的语种多达十二种。用户只要事先设定从何种源语言翻译至何种目标语言，然后选中 Word 环境内的任意中文或外文或中外文混合文，然后点选 Word 菜单中的工具→语言→翻译，机器就能自动将用户选中的那段文字从源语言翻译成所需要的目标语言，快捷并具有一定的可读精度。图 2 示出了笔者利用嵌入在 Word 软件中的 WorldLingo 翻译系统

将本篇文章的这一小段中文在大约十秒内翻译成英文的结果，从中可看出世界顶尖级软件供应商在研究开发多语种在线机助翻译上目前所达到的水平和效果。

在线机助翻译系统为人们阅读用别国语言文字书写的文字信息提供了重要手段，受到人们的青睐。但它的缺点仍然有二，一是每次都必须将所要阅读的文字信息粘贴到 WorldLingo 系统才能进行翻译，操作很不方便；二是所翻译出来的目标语言结果无法保持在原有的网页排版格式中阅读，这就大大影响了阅读网页的效果。基于此，激发了人们进一步研究开发称之为“网页在线即时翻译”的系统。

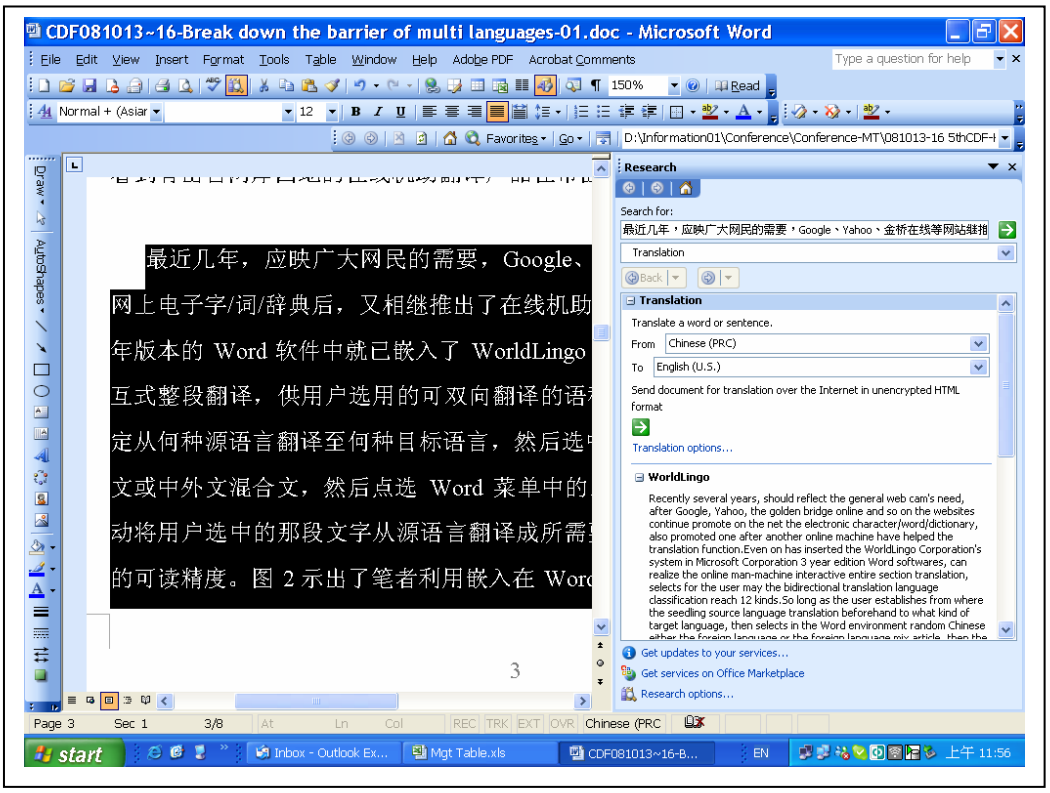


图 2. 用 Word 软件中的 WorldLingo 系统将一段中文在线翻译成英文的结果

3. 网页在线即时翻译

人们期望的阅读外文网页的理想化模式是，在存取世界各国各地的多媒体网页信息后能立即将整页文字信息快速翻译成所选定的目标语言，用目标语言替代原有的外文文字，但却继续保持原有网页的设计画面格式、动画不变，这就是本文定义的“网页在线即时翻译”的内涵。为了实现这一目标，目前在这一

方面代表世界最先进水平的微软和 WorldLingo 的作法是让用户调用他们研发的网页即时翻译工具“Instant Website Translator”或流览器翻译工具“Browser Translator”。图 3 示出了存取到的 Yahoo 一个动态英文网页画面，而图 4 示出了调用 WorldLingo 公司的 “Instant Website Translator” 经过大约二十秒钟在线即时翻译后的该网页画面。显然，翻译结果基本保持了原有网页的画面格式不变并具有相当的可读性，但同时也发现，原网页画面中的部分图片及文字在翻译后的画面上变成了空白。在操作上，也显得不方便不直接，用户必须把拟查询的网页地址粘贴到 WorldLingo 公司的 “Instant Website Translator” 中，系统才能存取那网页并开始翻译。

据介绍，流览器翻译工具 “Browser Translator” 的翻译功能同此类似，只是在存取某网站时一旦调用了流览器翻译工具，它就会把该网站的所有网页在后台逐一自动翻译到选定的目标语言，以方便用户查询而不必象调用“Instant Website Translator” 那样，每次翻译好一页网页必须回到 WorldLingo。至于其他公司的类似产品或工具或则效果雷同，或则水平更差。

图 3. 存取到的 Yahoo 一个动态英文网页画面

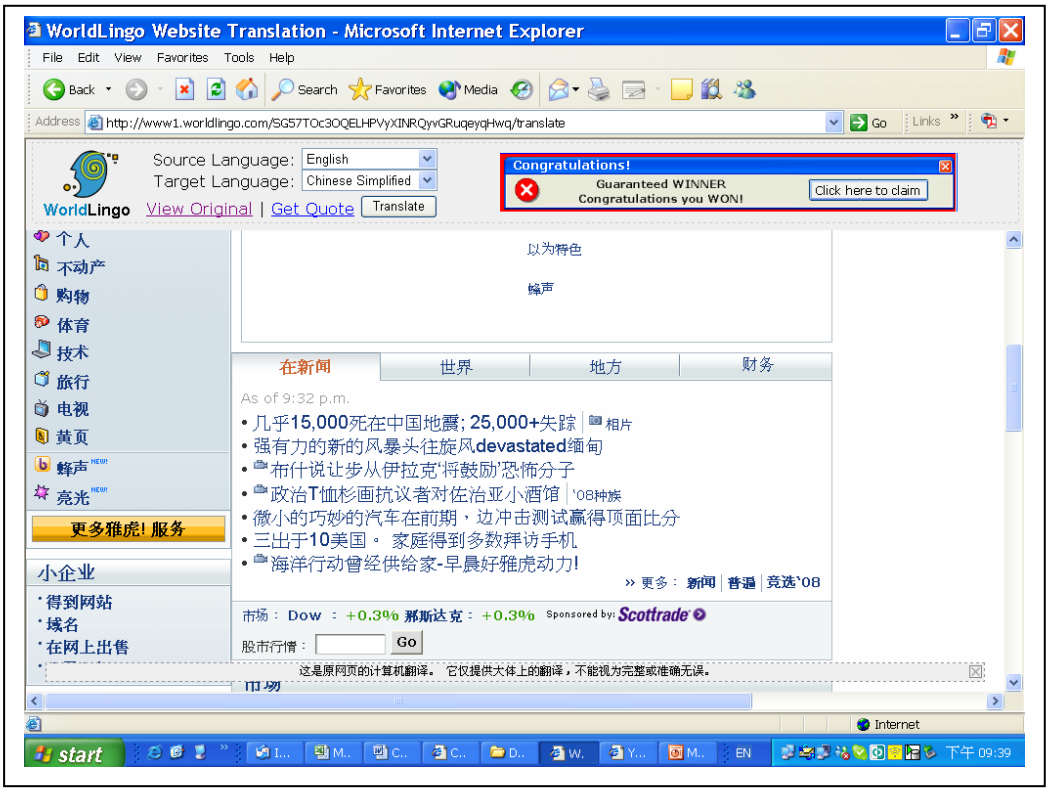


图 4. 用 WorldLingo 公司提供的工具经过大约二十秒钟在线即时翻译后的同一网页画面

与迄今为止微软、WorldLingo、金桥在线、Google 或 Yahoo 的网页在线即时翻译作法不同，我們的創新做法是将我们研制开发的网页在线即时翻译系统嵌入到 Internet 流览器内，在流览器的菜单上显示出供用户点击的“全文翻译”图标，用户只有在点击这一图标的时候，我们的在线即时翻译系统才投入工作，从而不影响流览器原有的一切功能。当用户存取并看到例如用葡萄牙文写就的网页信息后，点击一下“全文翻译”图标，于是启动网页在线即时翻译，就可让用户在原有的网页画面格式中阅读已经翻译成中文的全部信息了。图 5 示出了存取到的一个葡文网页画面，而图 6 示出了点击“全文翻译”图标经过大约二十秒钟网页在线即时翻译后的网页画面。显然，这种工作模式直接、快捷、方便，为那些不懂外语的普罗大众上网阅读外文信息，或让外国人阅读中文网页信息打开了极大的方便之门。

圖 5. 存取到的一个葡文网页画面

圖 6. 经过大约二十秒钟在线即时翻译后的同一网页画面

为了应对如何能正确翻译网页上可能出现的各种广泛内容，我们在网页在线即时翻译系统中混合采用了基于词库案例的翻译技术和基于上下文语义/词法/句法/语法分析的、基于形态和语义特征分析的机器自动翻译技术。最重要的是，采用了我们在长期研究开发中所总结出来的核心技术 TCT (Translation Corresponding Tree) 和 CSG (Constraints Synchronous Grammar) 。但正如开发全能专家系统注定是失败的情况一样，期望任何机器即时翻译系统能比较精确地翻译出包罗万象的网页内容那是根本不现实的。目前最高水平的机器翻译系统只能较好地在某一预定领域内达到百分之七、八十的翻译精度。为解决这一世界永恒难题，我们在机器翻译系统中进一步改进了机器自动学习功能并采用整篇网页分句强制培训的知识累积方式，以改进即时翻译系统的翻译精度。图 7 示出了某个任意存取到的葡文网页，图 8 示出了培训前对该网页在线即时翻译的结果，图 9 示出了经过多次培训后再对同一网页在线即时翻译的结果。显而易见，采用整篇网页分句人工培训的方式能让系统高效累积翻译知识，有效改进网页在线即时翻译的精度。当然，要真正让网页在线即时翻译能达到令人基本满意的程度，还有漫长的道路要走。然而，不走，那里就永远是一片荒漠空白；走的人多了，也就成了路！中文对外文的在线网上互译互通总不能靠外国人来代替我们去解决吧。

图 7. 某个任意存取到的葡文网页

图 8. 培训前对该任意存取到的葡文网页在线即时翻译的结果

图 9. 经多次培训后再对该同一葡文网页在线即时翻译的结果

結論

一如过去曾经说过的，只有对语言文字進行了正确无误的数字化处理，才能赋予电脑以更多的智能，使它变得更加強大好用，才能受到平民百姓的青睐，才能令其真正走入千家万戶。而只有到了那個時候，电子政府、电子商贸和电子事务才能真正得以实现，全民也才能真正享受到世界大同多元文化帶給全人

类的文明财富。从这个意义上讲，在国际互联网中消除多语言壁垒，实现中文同外文之间在线互译互通是实现这一目标的不可回避的重要任务，也是向语言文字数字化处理提出的新挑战，要求我们不僅对内要继续研究解决好兩岸四地的中文数字化互通問題，同時伴随着科学技术的飞速发展和人们对于无阻隔读懂国际互联网信息的新要求，也要求在中文数字化中大力研究开发中文同外文在线即时互译互通技术，把中文数字化处理技术提到新高度、新水平，為弘揚中华文字文化、同世界各国开展国际互联网更多的在线沟通和交流注入更多的活力。本文在這一領域的初步探讨期望能引发更多有识之士投入研究开发，迎來中文语言文字处理技术的更多成果，早日在国际互联网中消除多语言壁垒！

參考文獻

[1] Fai Wong, Mingchui Dong, and Dongcheng Hu. “Machine Translation Using Constraint-Based Synchronous Grammar”, Tsinghua Science and Technology of Tsinghua University, ISSN 1007-0214, Vol. 11 No. 3, 06/16 pp295~306, June 2006.

[2] Fai Wong, Mingchui Dong, and Dongcheng Hu. “Machine Translation Based on Translation Corresponding Tree Structure”, Tsinghua Science and Technology of Tsinghua University, ISSN 1007-0214, CODEN TSTEF7, Vol.11 No.1, 05/21 pp25-31, Feb. 2006.

[3] Isao Goto, Noriyoshi Uratani, Terumasa Ehara, Tadashi Kumano, and Hideki Tanaka. “A Multi-language Translation Example Browser”, The 9th Machine Translation Summit, pp.463-466, New Orleans, USA, 2003.

[4] Yilu Zhou, Jianlun Qin, Hsinchun Chen, and Jay F. Nunamaker “Multilingual Web Retrieval: An Experiment on a Multilingual Business Intelligence Portal”, Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS 38), Big Island, Hawaii, USA, January 3-6, 2005.

[5] Naoyuki Yoden, Kazuyoshi Harada, and Takeshi Yumura., “Personal Machine Translation Software for WWW Browser”, International Conference on Consumer Electronics, ICCE, IEEE vol. Conf. 15, pp. 234-235, New York, U.S. Jun. 5, 1996.