在 HNC 学术讨论会上的发言（2002 年，武汉）

旅欧见闻：国外 MT 和 HLT

冯志伟

（教育部语言文字应用研究所）

内容提要：本文介绍了作者在旅欧期间耳闻目睹的国外机器翻译和人类语言技术方面的情况，主要是德国、法国、英国、荷兰和美国的情况。

关键词：机器翻译，人类语言技术

MT & HLT Abroad

Feng Zhiwei

Institute of Applied Linguistics, Minister of Education

Abstract: The MT (Machine Translation) and HLT (Human Language Technology) in Germany, France, United Kingdom, Holland and USA were introduced in this paper.

Key Words: MT, HLT

1999 年到 2000 年，我应德国特里尔大学的邀请，在德国当了一年的客座教授（教授级别为 C3）。这是我第二次到这个大学担任客座教授。在德国期间，我又有机会到欧洲其他国家访问，对于欧洲的计算语言学研究有所了解，我还有机会遇见了一些美国学者，也了解到他们最近的研究工作。本文拟对这些情况做一介绍。

题目中的 MT 是英文 Machine Translation 的缩写，这个缩写大家都很熟悉了，众所周知，MT 是一种非常复杂而困难的技术。题目中的 HLT，可能很多人不知道是什么意思，这是近年来国外常用的一个新术语，是英文 Human Language Technology 的首字母缩写，它的意思是“人类语言技术”。这个缩写几乎成了“计算语言学”的同义术语，只是计算语言学更强调理论，而人类语言技术更强调技术。HLT 这个缩写术语最近常在国外文献中出现，因为在很多人看来，自然语言的计算机处理与其说是一门深奥难懂的学问，不如说是一种实实在在的技术。重技术轻理论，这是国内外共同的倾向，也是商品社会发展的要求。我认为这是一个好的倾向，我们搞理论研究的人，切不可轻视技术，我们不但在理论上要说得通，还要在技术做得好，让理论产生实用的效果，造福于人类。我想，这大概就是创造 HLT 这个新术语的学者的良苦用心。HNC 理论是开创性的理论，这个理论几乎涉及人类知识的各个部门，博大精深，但是，在实用技术方面目前还没有明显的成果。我希望 HNC 理论也同样要注意实际的技术，早日把理论付诸实用，让它开出实用的花，结出实用的果。

1. 德国

德国萨尔布吕肯大学(Saarbruecken Universitaet)是欧洲共同体 EUROTRA 计划的主要成员之一，在机器翻译方面卓有成就，现在，是德国人工智能学 DGKI (Deutsche Gesellschaft fuer Kunstliche Intelligenz)的挂靠单位，该大学的语言学系和计算机科学系都进行计算语言学的教学和研究。近年来主要从事文本自动生成（Automatic Text Generation）的研究。德国的机器翻译系统主要有 SUSY 和 Metal。SUSY 系统可以进行俄语、英语、法语、世界语的机器翻译。Metal 系统由西门子公司（Siemens）与美国 Austin 的德克萨斯大学（University of Texas, Austin）的语言研究中心（Linguistics Research Center,）联合研制，英德翻译系统已经商品化。现在由 Sail-Lab 公司继续开发，并开始研究汉语的自动处理。SAIL-LABS 购买了 Metal 的版权，开展多语言 MT 研究（包括汉语），中心在 Munich，分部设在荷兰、西班牙。

德国 LHT 的最著名的工作是 Verbmobile 计划，这个计划由卡尔斯鲁尔大学(Karlsruhe University)牵头，由德国联邦政府教育、科学、研究与技术部(BMBF)支持。其目的在于“通过工业及科学界尽可能多的分支领域的合作与集中，在下一个世纪的语言技术及其经济应用领域中为德国谋取国际领先地位”。Verbmobil 制定了 1993-2001 年的研制计划，其中自 1993 年至 1996 年的第一阶段计划吸收了德国、美国和日本的 32 个企业和高等学校的成员参加，政府投入资金 4690 万马克，企业投入资金 310 万马克，第一阶段的目标是建立非特定人的、面向会面安排交谈的口语语音翻译系统，其原型系统已经完成，将进一步进行实时自然语音翻译。最近，Verbmobile 计划的研究已经融入了 C-STAR。

C-STAR 是国际语音翻译联合会(Consortium for Speech Translation Advanced Research)的简称，1991 年成立，卡尔斯鲁尔大学是 C-STAR 最重要的成员。C-STAR 是一个以口语语音翻译为基本研究目标的国际合作组织，由来自 12 个国家的 20 个成员组成。核心成员有来自 7 个国家 7 个单位：美国的卡内基-梅隆大学(CMU)、日本的 ATR-ITL、德国的卡尔斯鲁尔大学 UKA (University Karlsruhe)、法国格勒诺布尔大学自动翻译研究中心 GETA-CLIPS、意大利的科学技术研究所 ITC-IRST、韩国的高级网络服务技术部 ETRI、中国科学院自动化研究所国家模式识别重点实验室(NLPR)。其他成员有德国西门子公司(Siemens)、香港科技大学等。C-STAR 把多种语言的口语直接翻译作为一个科学工程来进行，通过建立平台和演示来推动口语语音翻译技术的迅速发展，使 C-STAR 成为国际口语翻译技术转向工业应用的摇篮，以扫除人类的语言障碍。作为 C-STAR 核心成员的中国科学院自动化所国家模式识别重点实验室 NLPR(national Lab of Pattern Recognition)已经建立了口语翻译的试验系统的相关平台，正在开发可初步实用的汉英口语语音机器翻译系统。

2. 法国

我在 1978 年-1981 年曾经受中国科学技术大学研究生院的派遣，到法国格勒诺布尔大学(Universite de Grenoble)的应用数学研究所(Institut de Mathematique Applique de Grenoble，简称 IMAG)学习。IMAG 有一个遐迩闻名自动翻译研究中心，简称 GETA: (Groupe d’Etude pour la Traduction Automatique)，当时我的老师就是著名数学家沃古瓦教授(B. Vauquois)，他担任 GETA 的主任。沃古瓦教授桃李满天下，Prolog 语言的发明人 A. Colmerauer 就是沃古瓦的及门弟子。阔别了二十年之后又回到 GETA，我的老师沃古瓦教授已经去世，二十年以前的许多老朋友，如今已经成为国际计算语言学界的权威专家了，抚今思昔，人事沧桑，感慨万千。GETA 现在的主任是布瓦戴教授（Ch. Boitet）。他继承沃古瓦的事业，把 GETA 建设成国际著名的计算语言学研究中心，他把机器翻译专用软件 ARIANE-78 改进为 ARIANE-X，从而使得原来必须在 IBM 的 CMS 操作系统下运行的 ARIANE-78 程序可以在微型计算机环境下运行。ARIANE 包括 ATEF, ROBRA, TRANSF 和 SYGMOR 等软件包。其中，ATEF(Analysis de text en etat finit)是一个有限状态文本分析器，其数学模型是确定性的有限自动机(deterministic finite automata)，主要用于词法分析 （morphological analysis）；ROBRA (Arbre analysis)是一个树对树的转换器（tree to tree transformation），主要用于句法-语义自动分析（parsing）；TRANSF (TRANSFormation)是一个词汇转换器（lexical transformation），用于编制机器词典；SYGMOR (Systeme pour la Generation MORphologique)是一个非确定的有限自动机（non-deterministic finite automata），主要用于词法生成（morphological generation）。他们利用 ARIANE 软件进行多语言机器翻译（Multilingual MT）实验。目前也编制多语言机器词典。

我还访问了施乐公司（XEROX）欧洲研究中心 XRCE (XEROX Research Center in Europe)，这个中心就在离格勒诺布尔不远的麦兰（Meylan, Grenoble）。众所周知，施乐公司是专门制造复印机的，该公司的复印机是这样地有名，以至于在英文词典中，xerox 也获得

了复印机的意思，从专有名词变成了普通名词。

XRCE 也进行 LHT 的研究。他们花了十年的时间专门研制有限状态分析算法（Finite State Calculus）和隐马尔可夫模型 HMM (Hidden Markov Model)，设计了独立于具体自然语言的有限状态分析软件（Language independent software）。于连（Julian）博士用有限状态分析器建立了汉语文本的自动切分和标注系统（segmentation and tagging of Chinese text），效果良好。大家知道，有限状态自动机是基于乔姆斯基正则文法（3 型文法）的，一般只用来进行词法分析，在计算语言学中很不起眼。XRCE 用了十年的工夫，以达摩祖师“面壁十年图破壁”的精神，以小见大，做出了世界公认的成果，实在令人钦佩。安娜(Anna)博士用 HMM 研制的法语自动标注系统（Tagging based on HMM）也取得了相当精确的词类自动标注效果。XRCE 还建立了术语管理系统（Terminology management system），他们正在进行从双语语料库中自动抽出术语（automatic extraction of terms based on bilingual corpus）的试验。

我还访问了 XRCE 的多语言知识管理实验室 MLKM (Multi Languages Knowledge Management)。他们研制的自动翻译复印机可以在复印的同时进行关键词的英法翻译，输入的是英文，复印出来的是法文。他们还研制了具有自动文摘功能的复印机，输入的是整篇的文章，复印出来的是这篇文章的摘要。此外，他们还研制了数字化的名片(Digitalized name card)，可以在解码显示以数字方式储存在名片上的人的面孔；他们还研制了数字化的文本-图象扫描器，具有编辑功能。他们研制的用于互连网的全文搜索引擎，也给我留下了很深的印象。这些实用的 LHT，展示了计算语言学广阔的应用前景。

3. 英国

英国是个在计算语言学的理论研究方面卓有成就的国家。Gazdar 的广义短语结构语法(Generalized Phrase Structure Grammar，简称 GPSG)是计算语言学中影响最大的一种形式语法，得到广泛的使用。

与此同时，英国的语料库研究也取得丰硕成果。远在 1986 年，我在访问伯明翰大学(University Birmingham) 时就认识了著名的语料库语言学专家辛克莱（J. Sinclair）教授，当时他的 COBUILD 词典刚刚出版，他把出版社给他的第一本 COBUILD 词典送给了我，我有幸成为 COBUILD 词典的第一个读者。COBUILD 词典是完全根据英语语料库编辑的，所有的例句都来自英语语料库，我在读了这本词典后，曾经在《国外语言学》上写了一篇介绍辛克莱工作的文章，并且，同伦敦大学东方学院就建立汉语语料库的汉语标记集问题进行过非常仔细的讨论，后来又向当时的语用所领导提出过建立汉语语料库的建议，可是，孤掌难鸣，没有得到支持。这次我到欧洲，又先后两次遇到了老朋友辛克莱教授。一次是在意大利的蒙特卡梯尼（Montecatini），一次是在捷克的布拉底斯拉发（Bratislava）。J. Sinclair 教授向我介绍了他们工作的进展。一是他们针对不同的读者对象编写了 COBUILD 的不同版本，二是他们针对英语学习中的不同的问题，在 COBUILD 词典的基础上，编写了《英语语法大全》(Collins COBUILD English Grammar)以及英语语法系列的小册子。《英语语法大全》中的 1000 多个典型例句都来自 3 亿词的大规模英语语料库(The Bank of English)，系统地从语料库中归纳各种语法结构供读者灵活运用。英语语法系列小册子主要有：《介词》《构词法》《冠词》《易混淆词》《转述法》《同音异义词》《隐喻》《拼写法》《连词》《限定词及数量词》，例句都来自英语语料库。Collins 出版社在伯明翰大学建立了语料库语言学研究所，由托依拜特（W. Teubert）教授担任所长，专门从事语料库语言学的研究。辛克莱教授现已退休，托依拜特继承他的工作，而辛克莱教授则迁居到意大利托斯坎地区，建立托斯坎词中心（TOSCAN Word Center），该中心定期举办语料库语言学培训班，面向全世界招生。

近年来词汇语义学(lexical semantics)的研究很热烈。我在希腊雅典召开的 LREC（Language Resources and Evaluation Conference）会议上，遇见了英国布来顿大学（Brighton

University）的基尔加里夫（A. Kilgarriff）博士，他发起了一个叫做 SENSEVAL 讨论会，通过网络讨论多义词的自动歧义消解问题，吸引了世界各国词汇语义学研究者来讨论这个计算语言学中的难题，许多青年研究者参加讨论，有兴趣的读者可以和他们联系(Brighton University, Dr. Adam Kilgarriff, email: mpalmer@linc.cis.unpenn.edu)。

4. 荷兰

欧洲共同体在 1982 年开始实施 EUROTRA 计划的同时，还支持荷兰 DLT 多语言机器翻译系统的可行性研究。从 1984 年开始，DLT 改由荷兰政府和荷兰的一家软件公司 BSO 各出资一半对此系统的研制进行长期的支持，从 1984 年到 1992 年每年投资均在 100 万美元左右。DLT 系统原打算九十年代中期开始实用化，可是至今尚未得到满意的结果。

5. 欧洲的计算语言学学术团体

5.1 跨欧洲语言资源基础建设学会 (Trans-European Language Resources Infrastructure，简称 TELRI)：负责人是托依拜特（W. Teubert）教授，原来挂靠在德国曼海姆的德语研究所(IDS, Institut fuer Deutsche Sprache)，现在挂靠在英国的伯明翰大学语料库语言学研究所，主要任务是协调欧洲的多语言信息处理和多语言语料库的建设。建立了语言资源和工具库 TRACTOR (Research Archive of Computational Tools and Resources)，注册后可以访问。目前 TELRI 致力于建立 EUROVOCA，这是欧洲语言的词汇库。2000 年 9 月 22－26 日在斯洛文尼亚的首都卢布里雅那（Lubljana, Slovenia）召开了学术讨论会，主题是知识的自动抽取（Automatic knowledge extraction）。我参加了这次会议并在会上发言。

5.2 欧洲语言资源学会(European Language Resources Association, 简称 ELRA)：负责人是意大利比萨大学计算语言学研究所的扎波里（Zampolli）教授。ELRA 的任务是搜集、传播语言资源并使之商品化，对于语言资源的使用提供法律支持。下设欧洲语言资源传布处 (European Language resources Distribution Agency，简称 ELDA)：负责研制并推行 ELRA 的战略和计划。ELRA 组织语言资源与评价国际会议(Language Resources & Evaluation Congress, 简称 LREC)，每两年一次。第一次会议于 1998 年在西班牙的 Grenade 举行。第二次会议于 2000 年 5 月 31 日至 6 月 2 日在希腊首都雅典(Athens, Greece)举行。我参加这次会议，担任大会的国际顾问委员会委员，并应大会主席扎波里教授的邀请，在会上作了专题报告。

5.3 国际术语学和应用语言学中心(Centrum international fuer Terminologie und angewendte Linguistik, 简称 CiTaL)：负责人是德国康斯坦茨高等技术学院（Fachhochschule Konstanz）的托马森(W. Thomassen)教授，主要从事术语数据库和机器翻译的研究。去年托马森教授退休，新任的负责人是斯洛维格(P. Slowig)教授。我是 CiTaL 的技术顾问。

6. 美国

在欧洲期间的一些学术活动中，我遇到了美国的一些同行。例如，在捷克的布拉底斯拉发(Bratislava)举行的 TELRI 会议上，我见到了格语法的创始人菲尔摩(C. Fillmore)，在希腊的雅典举行的 LREC 会议上，我见到了美国宾州树库(Penn Tree)的负责人尤喜（A. Joshi）教授。在欧洲我还可以及时地阅读到各种外文学术刊物，眼界大开。在同美国这两位著名学者的交谈中，在阅读学术刊物中，我了解到美国 MT 与 LHT 的一些情况，下面，我主要介绍一下美国 IBM 公司基于统计的机器翻译研究。

1994 年，美国 IBM 的 Adam L. Berger 等人，采用全统计的法英语料库对齐方法，经过

五年的努力，利用对齐了的法语和英语的 2205733 个句子对，在 IBM 15 台 530H Power 工作站上，运行 3600 小时，开发了一个法语-英语的机器翻译系统。

与基于规则的 SYSTRAN 系统相比较（不受限文本），这个基于统计的机器翻译系统的成绩是令人鼓舞的：

	译文流畅的句子	译文可读的句子
SYSTRAN .	.540	.743
IBM (自动翻译)	.580	.670

可以看出，译文流畅句子的比例，IBM 的系统超过了 SYSTRAN，而译文可读句子的比例，SYSTRAN 略高于 IBM 系统，这说明，基于统计的机器翻译系统有可能超过基于规则的机器翻译系统。

IBM 的基于统计的机器翻译的特点是：

- 不以词汇作为处理单元，而以语段 (chunk) 作为处理单元(Abney, 1995);
- 采用相似理论和对齐方法，使用统计方法进行句法处理。

这种基于统计的机器翻译方法，主要工作是：

- 构造大容量的网络上的双语电子基本词典；
- 严格定义扩展语段，给出形式定义，利用有限状态自动机进行原语语段(chunk)的自动获取；
- 语段的分类与归纳，构造模板库(Template Base);
- 构造大容量的网络上的双语语段词典(chunk Dictionary), 约 1000 万条；
- 研究机器可读电子词典中搭配短语的获取算法，构造大容量的网络上的双语搭配词典；
- 研究模板的相似度算法以及双语对齐(alignment)的方法；
- 完善网络上的双语专业术语词典，给出基于词尾变化的语料库短语词性标注算法；
- 研究原语未登录词的识别算法；
- 建立基于语段的机器翻译系统。

IBM 公司的西拉魏格纳（F. Ciravegna） 提出了确定英语语段的原则，英语语段形式是有严格限制的，按层次可以分为如下 4 种：

1. 单义词
2. NP = (adv* adj*) noun* (adv* adj*)
   VG = (adv)* Verb* (adv)*
3. DP = Det* NP
   PP = Prep Det* NP
   VP = (adv)* (auxiliary)* VG
4. 常用子句

IBM 机器翻译系统的双语语段词典是机器自动从语料库中获取的，共 1000 万条。这种大规模的语段词典，成为了机器翻译最重要的知识源。

美国菲尔摩(C. Fillmore)教授向我介绍了他最近主持的框架网络 FrameNet 的进展情况，这个课题得到了美国国家基金的支持，采用框架来描述英语动词的语义关系，有很大的实用价值和理论意义。

美国尤喜(A. Joshi)教授向我介绍了 LDC (Linguistic data Consortium)的情况，LDC 是语言数据资源联合会，会员把自己的语料库提供给 LDC，同时就可以共享 LDC 的资源，现在 LDC 现在已经:有 163 个语料库（包括文本语料库和口语语料库），他们还有中文的树库。语料库的建设是很艰巨的工作，为了避免重复劳动，共享语料库的资源，这是很聪明的办法。

现在我国已经建立了不少的汉语语料库，如何实现资源共享，避免重复劳动，LDC 的经验是值得我们借鉴的。

我在欧洲一年，见闻有限，上述介绍难免挂一漏万，仅供参考。不妥的地方，请诸位批评指正！

HNC 在理论建设方面的成绩是引人注目的，但是，在实用技术方面，特别是在 MT 和 HLT 方面还显得比较薄弱，希望 HNC 在不久的将来，在 MT 和 HLT 方面都取得卓有成效的突破，从而显示 HNC 理论的威力。