

·实践平台·

# 机器翻译在 CADAL 中的应用

黄 晨 陈海英 ( 浙江大学图书馆 浙江杭州 310027)

摘 要: 文章将对在“中美百万册数字图书馆(简称 CADAL)”项目中提供双语服务所涉及的机器翻译技术以及如何将该技术在数字图书馆建设中应用进行讨论并提出解决方法。

关键词: CADAL 机器翻译 数字图书馆

中图分类号: H085

文献标识码: A

文章编号: 1003-6938(2007)01-0066-03

## Application of Machine Translation in China- America Digital Academic Library

Huang Chen Chen Haiying ( Zhejiang University Library ,Hangzhou ,Zhejiang ,310027)

Abstract : This paper briefly introduces the main ideas of machine translation (MT) techniques and then discusses how to use them in China- America Digital Academic Library(CADAL) for translating source language to the target language.

Key words : CADAL ; machine translation ; digital library

CLC number : H085

Document code : A

Article ID : 1003-6938(2007)01-0066-03

### 1 引言

中美百万册书数字化工程是由中美双方科学家共同发起的旨在建设包含 100 万册图书的数字图书馆研究与开发项目(<http://www.cadal.cn>),被列为教育部“十五”重点项目之一——“中英文图书数字化国际合作计划”(China-America Digital Academic Library ,CADAL)。<sup>[1]</sup>该项目基于开放框架结构,其目标是建设面向教育和科研的百万册(中文 50 万册,英文 50 万册)规模的图书数字化文献资源,为教学科研提供强有力的数字资源支持,推动图书数字化资源的共享。这将是迄今为止全球范围最大的数字图书资源库(百万量级)。

数字图书馆的最终目的是信息服务,今天的用户已不再满足于通过网络检索出一大堆原始信息,而是要求一个系统集成、媒体丰富、深度加工的信息,甚至是一个包含知识和解决方案的信息服务。<sup>[2]</sup>因此,CADAL 项目不仅以数字图书代替印刷图书为读者服务,而且对这些资源进行深加工,挖掘

蕴涵在资源库中的知识,增加服务方式,拓展服务范围,使数字资源能够更有效地为广大读者服务。其中利用机器翻译(Machine Translation ,MT)技术,提供双语乃至多语种服务,就是 CADAL 项目准备采取的一项特色技术服务。

#### 1.1 国内外的的发展

机器翻译是自然语言理解(Natural Language Understanding ,NLU)中最早的一个研究分支,它是利用计算机把一种自然语言转变成另一种自然语言的过程。用以完成这一过程的软件叫做机器翻译系统。

1949 年,美国 Rockefeller 基金会自然科学部门的负责人 Warren Weaver 发表了一份以《翻译》为题的备忘录,正式提出了机器翻译问题。80 年代中期以后,无论是经典语言学理论还是新兴的科学——计算语言学理论的发展都日益完善,出现了不少商品化的系统,如美国的 SYSTRAN 系统,美国 Texas 大学与西德 Simon 公司合作研制的 METAL 系统,日本日立公司的 ATLAS 系统及法国 Grenoble 大学的 CETA 系统等等。

基金项目: 本文系教育部“高等学校中英文图书数字化工程(CADAL)”支持项目系列研究论文之一。

收稿日期 2006-02-27,责任编辑 陈笑悦

我国机器翻译的研究从一开始就得到了国家的高度重视。早在 1956 年它便以“机器翻译/自然语言的数学理论”列入了当时的《科学发展纲要》。80 年代中期到 90 年代初期产生了两个在中国机译史上具有重要意义的实用化系统,分别是军事科学院研制的“KY-1”英汉机译系统和中科院计算所研制的“863-IMT”英汉机译系统。

近年来的机译系统大体上有这样一些特点:多数配有大规模的多领域的专业词典,能在网上运行,有相当不错的方便用户的界面。新的应用领域的机器翻译研究,如对话翻译系统的研发等也已开始。<sup>[3]</sup>

## 2 机器翻译理论与技术

### 2.1 机器翻译方法论

传统的机器翻译体系基本可以纳入基于知识的方法(Knowledge Based MT, KBMT)范畴,也称为基于规则(Rule Based)的方法。KBMT 面临的最大的问题是需要海量的计算语言学资源,如大规模的句法和词汇系统。从目前语言知识工程的进展来看,为一个通用的、高质量的机器翻译系统手工构建这些资源在可以预见的将来仍然是不现实的。

由于 KBMT 面临的困境与挑战,1984 年日本东京大学的长尾真教授提出了基于类比的机器翻译方法。<sup>[4]</sup>在这篇著名的论文中,长尾真主张,语言学数据是比语言学理论更可靠的知识源,因此也可以为机器翻译系统奠定更坚实的基础。他建议使用无标注的实例数据库和一个等价词对的集合作为系统的知识源,翻译引擎主要负责计算输入句子和候选实例中词汇间语义的相似性。

很多研究者对长尾真的方法进行了扩展,现在统称为基于实例的机器翻译(Example Based MT, EBMT),即从数量日益增长的机器可读文本出发,使用经验主义的方法构造自动翻译过程所需的语言知识。与 KBMT 中提倡的尽量应用深层语言学知识的主张相反,经验主义机器翻译方法中这些资源通常是相对浅层的模板化表示,有的甚至就是表层的词汇统计信息。

另外一种经验主义的 MT 系统称为随机机器翻译系统(Statistical MT, SMT)。以 Hansards 英法双语语料为基础,IBM 的 Brown 等人实现了第一个 SBMT 系统模型——Candide<sup>7</sup>。这种翻译方法是将翻译系统看作是一个噪音信道,令表示以  $e$  为输入通过信道获得译文  $f$  的概率,则给定输入  $f$ , 机器翻译的过程可以描述为在目标空间  $E$  中寻找满足条件的句子,作为翻译系统的输出结果。

尽管这种随机机器翻译模型解决了知识的获取问题,但其模型巨大的参数空间以及由此而需要的数据资源和计算资

源都是十分可观的。Brown 等人的初步实验结果表明,基于这种“纯粹”统计的方法仅获得不到 40% 的准确率,而附加了基本词法信息后,其准确率可提高到 60%。这似乎预示着将 SBMT 与 KBMT 相结合才是未来研究中真正的出路。

### 2.2 机器翻译过程

机器翻译过程主要由源语分析和目标语生成两部分组成。源语分析是所有现代机器翻译系统的基础,翻译的质量本质上依赖于分析的质量和深度。所谓源语分析,就是遵循一定的语言学基础,寻求源语文本的表示形式与其对应内容之间所存在的映射关系的过程。典型的源语分析手段为:依据与源语文本所表达含义相关的词汇、句法结构、单词和句子的顺序,灵活地找出目标语译文。源语分析涉及多个不同层次,按复杂度递增顺序可划分为以下几个阶段:

(1)形态分析:用于获取源语言词汇原形。在机译系统的研制中,两层分析法是普遍采用的形态分析理论,<sup>[5]</sup>而有时也采用不太通用但更适合于特定语言、特定任务的方法。

(2)句法分析:用于摘取源语文本短语结构、句法结构的依存性,即确定输入文本中词汇的词性、短语边界及短语的内部结构。

(3)语义分析:利用文本含义描述语言建立知识结构,反映源语文本的词汇、词义及相互之间所存在的语义依存关系,可消除词义歧义、修饰歧义、分词歧义等等。

(4)语用分析:根据源语文本元素之间所存在的各种面向应用领域和修辞的关系,建立源语文本语义结构。语用分析主要解决指代歧义问题、通过语义格角色约束的确定、比喻和换喻的理解、<sup>[6]</sup> 坏结构输入所引起的问题以及省略情形<sup>[7]</sup>等等。

与源语分析对应的是目标语生成,它可以看作是源语分析的逆过程,主要完成以下两项任务:

(1)文本规划:对各种表达方式进行选择,确定欲实现的目标语文本的有关内容、修辞方式等信息。

(2)表层实现:根据目标语语法,将由词汇组成的句法表达式映射为表层字符串。

## 3 CADAL 中的机器翻译实践

### 3.1 目标与任务

CADAL 项目将完成的百万册数字图书中,中英文图书各占一半,为了使全球用户更有效地使用这些图书资源,项目拟覆盖 1 万册中英文双版本图书,并在此基础上,利用机器翻译技术支持双语服务。具体任务为:

(1)书名、作者等重要信息由人工翻译,或先由机器翻译,也可采用各种辅助翻译系统,再加以人工校对。

(2)机器翻译作为系统的基本组成部分,提供不同级别

的即时服务。这种服务支持对有 XML 标记内容的翻译。

(3) 机器翻译需要与其他服务相结合,如跨语言检索、专名检索等。翻译服务可以是分布式的,也可以是集中式的,但本质上是分布式的。

(4) 以 Web Service 为构建高层分布式应用的接口。为了高效应用,同时提供操作系统级的接口。

### 3.2 系统评测

根据 CADAL 有关子项目的设计,系统的构建基础建立于对现有机译系统的评测。在明确翻译内容和百万册书文本及相关元信息的基础上,评测国内外现有的翻译系统,选择较好的多家机器翻译系统作为集成对象。

为此,我们参考了由美国国家标准和技术研究所(NIST)举办的机器翻译评测结果。2002年6月,包括IBM公司,Carnegie Mellon大学,南加州信息科学研究所(USC/ISI),德国亚琛(RWTH Aachen)大学,微软研究院(Redmond)和中国科学院计算研究所在内的6家研究机构的机器翻译系统,参加了NIST的首次正式评测。同时,NIST还评测了SYSTRAN公司的商用机器翻译系统作为一个横向比较。<sup>[8]</sup>

评测结果显示德国亚琛大学,CMU大学和ISI的机器翻译系统性能优越,接近甚至超过了SYSTRAN公司的商用翻译系统。亚琛大学采用的是随机机器翻译(SBMT)模型,将传统的噪声信道翻译模型改进为最大熵模型,并且把基于词的对齐模型增强为基于短语的对齐模型,大大改善了翻译质量。CMU大学的Mega2RADD翻译系统通过翻译引擎,把基于短语翻译的随机机器翻译(SBMT)系统和基于实例的翻译(EBMT)系统集成为一体,通过比较和选择输出最优的翻译结果。ISI研究所开发的Re2Write翻译系统采用IBM-4统计模型为原型,加入了语法分析模块和联合短语翻译模块(KBMT),也有效地提升了系统的翻译质量。这三个系统所采用的模型与技术改良,给我们在CADAL应用中的实践提供了很好的启示:采用单一的翻译策略,不管是基于规则,还是基于语料库,都只能解决一部分问题。同一个系统里融合基于规则和基于语料库的多种机器翻译方法和技术,是这三个系统的一致特点。

### 3.3 技术模型

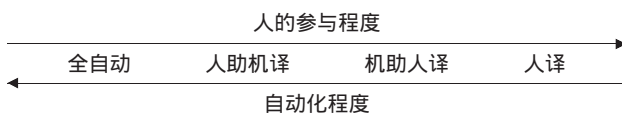
上述的评测与当前的研究使我们确定,MT研究的真正进展来自混合方法(Hybrid Approaches),也就是将多种MT方法集成在一个统一的MT环境中,形成多引擎MT系统。这与CADAL项目可行性研究报告的预计也是一致的。

我们设想,通过与本项目的美方合作单位卡内基·梅隆大学(CMU)的协同工作,以CMU现有的Mega2RADD翻译系统为基础,吸收亚琛大学的思想,将噪声信道翻译模型改

进为最大熵模型,模型的参数估计可以利用大规模文本样例中得到的大量统计数据。其中先验概率 $P(E)$ 可以通过构造合适的英语语言模型加以估计,而信道概率 $P(F|E)$ 则可以从由一个计算源文本部分相对应目标文本部分的自动过程建立了联结(Alignment)的并行文本中进行估计。为此,CADAL中的MT系统除了一般的计算机系统都有的硬件和软件(自然语言的语句分析与语句生成程序)外,还需有一个特别的组成部分,即语言知识库,包括静态的词典、语法规则库等,也包括动态的上下文相关信息。

另一方面,在集成多个引擎的思路下,将采用公开的MTSDK(<http://lan.cpip.net.cn/>)作为新规范的基础,构建不同级别的翻译服务。对不同目标(如速度和准确性之间的折衷)采用不同的引擎。例如,对作者、标题、句子、段落、摘要、全文的翻译可能需要不同的翻译机制,允许翻译服务的请求者对翻译内容做出某种标记。当然,翻译引擎必须能够理解这种标记。

第三点,根据人在MT中起的作用,可以用下图来表示不同的方式。



从全自动翻译到人工翻译,CADAL项目重点关注于“人助机译”和“机助人译”。因此在分级翻译机制下,进行适度的人工介入,也是CADAL项目提高翻译质量的手段之一。

## 4 结论

CADAL项目将充分采纳不同系统的技术优势,制定开发目标,建设自己的机器翻译服务。系统采用多种翻译处理策略,包括规则分析、类比推理、统计分析,用面向对象的多类型数据库来管理翻译所需的各种信息,并且提供人机交互接口,实现人工对翻译结果的干预。为了获得KBMT所需要的计算语言学资源,参考语义网(Semantic Web)的研究成果,引用语言本体(Ontology)实例构建词汇系统,<sup>[9]</sup>也是CADAL值得关注的领域。

作为一个全球共享的数字图书馆,CADAL将利用信息技术对各种层次的用户提供不同形式的服务,逐步保证所有人群都能够有效地利用数字资源进行学习和工作。<sup>[10]</sup>随着计算机应用技术的不断发展,我们不仅可以利用机器翻译技术把文本信息转换为用户熟悉的语种,还可以利用机器翻译技术进行基于语义的跨语种信息检索。(下转第85页)

一下用户如果按照自己的建议去做利益何在,要了解客户的心理和他所咨询的问题,要帮助客户权衡利弊得失,讲清利害关系,看准客户的需求,说服才能有的放矢,取得成效。

### 3.6 发展的同盟与共赢

咨询客户无论是政府还是企业,都是由很多人组成的“生命体”,是一群人为了满足另一群人的需求而主动创造世界的组织机构。一个优秀的咨询机构既要能深入其组织机体内部悟它的内动力(主流的)与潜意识(积极的),又要能站在现实环境与未来发展的角度上,从独立的、客观的角度观察分析问题,从而能够发现一些客户自己忽略的或有意忽视的东西。咨询人员找出客户最关注的价值领域,分析与客户建立起的联系,了解谁是主要竞争对手以及他们是如何吸引顾客的。咨询人员要提高市场反应速度,建立预测系统,为客户提供有价值的服务,不断提高知名度和美誉度。

咨询机构应该将咨询客户视为双赢的同盟,而不仅仅是推销咨询服务的对象。咨询机构与咨询客户之间实际上是处于同一条价值链上的实体,一荣俱荣,一损俱损。咨询机构要时时刻刻把咨询服务对客户价值放在第一位来看,只有实现了客户感知价值的最大化,才能实现咨询机构价值的最大化。这样咨询机构与客户在价值链中获得的利润同时达到最大,从而建立一种长期的同盟与共赢的关系。

#### 参考文献:

- [1] 郝沐平.咨询用户类型、心理及需求特点研究[J].中国图

书馆学报,1999,(5):78-82.

- [2] 张兵.也谈顾客满意的咨询师[J].中国质量认证,2002,(6):46.
- [3] 吴贺新,张旭.现代咨询理论与实践[M].北京:科学技术文献出版社,2000:77-90.
- [4] 陈翔宇,甘利人,郎诵真.现代咨询理论与实践[M].成都:电子科技大学出版社,1994:63-64.
- [5] 余明阳.咨询学[M].上海:复旦大学出版社,2005:77-90.
- [6] 常桦.咨询师手册[M].北京:中国纺织出版社,2005:84-85.
- [7] (英)菲利浦·萨德瑞著;段盛华译.管理咨询:优绩通鉴[M].北京:中国标准出版社;香港:科文出版有限公司,2001:94-95.
- [8] (美)伊莱恩·比斯著;孙韵译.咨询业:基础与超越[M].北京:机械工业出版社,2002:3-4.
- [9] 张承耀.寻求咨询:企业借助外脑的重要方面[J].镇江学刊,2004,(3):31-32.
- [10] 田同生.客户关系管理的中国之路[M].北京:机械工业出版社,2001:5-7.
- [11] 陈涛,孔庆杰.信息时代咨询业的定位与发展策略分析[J].图书馆学研究,2005,(1):93-95.

作者简介:张怀涛(1957-),男,中原工学院图书馆研究馆员、郑州大学信息管理系硕士研究生导师。

(上接第68页)这对于知识的有效利用无疑是具有划时代意义的尝试。

#### 参考文献:

- [1] 发改委.中英文图书数字化国际合作计划[Z].2004-1649号
- [2] 侯雅楠.网络环境下的知识挖掘[J].情报科学,2003,(8),887-890.
- [3] 董振东.中国机器翻译的世纪回顾[J].中国计算机世界,2000,(1).
- [4] Nagao M. A framework of a mechanical translation between Japanese and English by analogy principle [A]. In: Elithorn A and Banerji R eds. Artificial and Human Intelligence, Edited Review Papers presented at the International NATO Symposium [Q]. Amsterdam: NATO Publications, 1984: 173-180.
- [5] Karttunen L. KIMMO: A Two-Level Morphological Analyzer [J]. Texas Linguistic Forum, 1993(22):65-186.

- [6] Fass D. and Wilks Y. Preference Semantics, III - Formedness and Metaphor [J]. Computational Linguistics, 1983, 9(2):178-187.

- [7] Carberry S A Pragmatics-based Approach to Understanding Intersentential Ellipsis [A]. In: Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics [Q]. 1985: 188-197.
- [8] BL EU: a method for automatic evaluation of machine translation [Q]. Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, 232-240, ACL 2002.
- [9] 黄晨.信息服务的趋势: Semantic WEB [J]. 图书馆杂志, 2004,(3):55-57.
- [10] 陈海英,黄晨.在 CADAL 中应用文语转换技术 [J]. 情报学报, 2004,(5).

作者简介:黄晨(1970-),男,硕士,浙江大学图书馆数字资源建设中心副研究馆员;陈海英(1963-),女,学士,浙江大学图书馆数字资源建设中心副研究馆员。