

蒙古语的机器翻译系统

牧仁高娃

(内蒙古大学 蒙古学学院, 内蒙古 呼和浩特 010021)

摘要: 本文探讨了机器翻译技术的历史变革和实现方法, 并简单介绍了目前的几个蒙古语机器翻译系统。

关键词: 机器翻译; 机器翻译方法; 蒙古语的机器翻译

中图分类号: H212

文献标识码: A

一、引言

突飞猛进的计算机科学技术发展带动了各学科领域的发展和社会进步, 也不例外给机器翻译曲折发展带来了曙光。

机器翻译(Machine Translation)是使用计算机把一种语言(源语言/source language)翻译成另一种语言(目标语言/target language)的一门新学科, 同时也是一种新技术。它涉及到语言学、计算机科学、数学等许多学科, 是非常典型的多边缘的交叉学科。^①

国内外大型语言(例如英语、汉语)的机器翻译研究开始较早, 技术也逐渐趋于成熟。与此相比, 蒙古语的机器翻译研究起步较晚, 从上个世纪末到现在, 只有近十年的历史。在这期间虽然取得了一些可喜的成就, 但还是存在着很多问题, 如面向语言信息处理的基础建设不够充分、发展不平衡横和滞后等等。在这信息时代, 若某种语言不想被淘汰, 而要力求发展, 那就必须纵观走过的历程, 总结已有的经验, 为下一步发展做基础是很有必要的。

本文根据已搜集到的蒙古语机器翻译系统的相关资料数据, 简单介绍了几个典型的蒙古语机器翻译系统, 及其所用的方法和遇到的问题。

二、机器翻译的发展

20 世纪 30 年代初, 法国科学家 G·B·阿尔楚尼提出机器翻译的设想。时至今日, 机器翻译经历了“两次高潮, 两次低落”。计算机的发展水平、人们对自然语言理解的认识水平和人们对机器翻译的需求水平是机器翻译发展过程中的三个决定性因素。

20 世纪 50 年代初, 计算机的出现和冷战时期情报翻译的需求促成了机器翻译历史上的第一次研究热潮, 英国工程师布斯(A. D. Booth)和美国工程师韦弗(W. Weaver)在讨论电子计算机的应用范围时, 就提出了利用计算机进行语言自动翻译的想法。在学术界的积极倡导和企业的大力支持下, 机器翻译研究一时兴盛起来。1954 年, 美国乔治敦大学在国际商用机器公司(IBM 公司)的协同下, 进行了世界上第一次机器翻译试验, 该系统把 50 个俄语句子翻译成英语, 无需译后编辑, 翻译质量非常好。接着, 苏联、英国、日本也进行了机器翻译试验, 机器翻译出现了热潮。早期的机器翻译大多受到韦弗的影响, 把机器翻译看成一种机械的解读密码的过程, 而没有意识到机器翻译在词法分析、句法分析以及语义分析等方面的复杂性。由于对自然语言认识水平低, 机器翻译的质量很差。1966 年, 在 ALPAC 报告的影响下, 机器翻译研究走向低潮。

从 60 年代中期到 80 年代初, 计算语言学和人工智能获得了长足进步, 人们在利用计算机处理自然语言方面有了很大的进步, 这为下一个高潮期的到来奠定了基础。作为智能信息处理的一个

重要领域，机器翻译重新受到了各国的重视。许多发达国家相继投入了巨额资金研究和开发机器翻译，形成机器翻译的第二次高潮。欧共体的 EURORA 计划和 DLT 系统、日本的 Mu 系统和 ODA 计划、美国 SYSTRAN 的机器翻译研究等，都是这时期著名的机器翻译研究项目。在重视语言知识的表示机制的同时，人们开始着手机器翻译策略的研究。除了传统的基于规则的翻译方法，基于实例的机器翻译、基于统计的机器翻译等新的机器翻译方法不断涌现。但到 90 年代初，再一次因翻译效果不如人意而落入低潮。

从上世纪 90 年代中期以来，随着互联网的迅猛发展和经济全球化时代的到来，人们对机器翻译软件的需求日益增长。②1990 年由 IBM 公司的 P. Brown 等人提出的基于信源信道思想的统计翻译模型引起了人们的重视。随着研究者的不断努力，出现了基于短语的统计翻译模型、基于最大熵法统计机器翻译模型等和与其相应的机器翻译系统。

三、机器翻译的实现方法

机器翻译技术从总体上也可以分为基于规则的和基于语料库的两大类。

（一）基于规则的方法

基于规则的机器翻译方法（RBMT）是模拟人脑的翻译处理过程，即先用规则对原文进行分析、理解，弄清原文表达的意思，生成译文。这与原文表达的意义与分析的深度有关，虽然分析越深，对原文表达内容的把握也就越丰富，生成的质量就越高。但随着分析深度和难度增强，分析结果的可靠度就会下降。在对自然语言的认识尚未取得突破之前，基于规则的 MT 方法能得到很好的结果的可能性是很小。

但在嵌入式机器翻译中基于规则的 MT 方法是比较经济的。如现在的“快译通”等电子辅助学习机产品内的句子翻译的实现都是基于规则的 MT 方法。

基于规则的方法可以分为直接词对词的翻译、基于转换的翻译和基于中间语言的翻译。

（二）基于语料库的方法

基于语料库的方法从上个世纪 80 年代中后期开始发展起来，也被称为经验主义方法。基于大量的真实文本，建立大规模的语料库，可以给 MT 研究提供各种语言事实和语法规律。它可以分为基于统计的方法和基于实例的方法。基于统计的方法的主要内容是双语句对的对齐，通过词汇同现的可能性来计算一种语言的一个词映像到另一种语言的一个词的概率。基于实例的方法的基本思想是对于输入的任何一个句子 S，通过一定的评分机制在语料库中匹配一个最相近的句子 S1，这样 S1 在语料库中的译文 T1 就可以作为 S 的译文。

（三）混合策略

现在越来越多的机器翻译系统用混合策略。混合策略是把基于规则的方法和基于语料库的方法结合起来使用。实践证明，这种方法效果很好。例如有由基于实例的翻译方法和基于规则的方法相互结合而成的基于范本的翻译方法等。

四、蒙古语的机器翻译

中国少数民族语言的机器翻译研究开始比较晚，在经济、技术和研究人员方面都相对短缺。近年来蒙古语、藏语、维语等少数民族语言的机器翻译研究工作相继开始进行，大多数都在研究和实验阶段，几乎尚未到真正的实际应用阶段。由于蒙古语自身的形态变化多样化等特征对蒙古语的机器翻译研究也带来了相应的难度。下面简单介绍几种典型的蒙古语机器翻译系统。

（一）面向政府文献的汉-蒙机器翻译系统

2001 年，由内蒙古大学蒙古学研究院、中国科学院计算机研究所、北京大学计算机语言研究所共同合作开发的“面向政府文献的汉蒙机器翻译系统”通过国家验收。这标志着蒙古语文翻译事业

掀开了新的一页。

“面向政府文献的汉蒙机器翻译系统”是国家 863 计划的重要项目。旨在研究开发出一个面向政府文献的（如政府工作报告、各种决议等）的汉蒙机器翻译系统，使政府文件的少数民族翻译走上现代化、高效率的道路。

该系统是基于规则的翻译系统。研究包括汉语分析、蒙古语生成、翻译、用户接口四个组成部分。其中，汉语分析部分包括对文本进行分词、标注、分析等一系列处理工作。蒙古语生成部分由蒙古语的语言模型、蒙古语的生成规则集、从汉语到蒙古语的转换规则集和汉蒙对照机器词典组成。翻译部分中由一个汉蒙对照政府文献语料库，一部汉蒙对照机器词典和汉蒙机译软件三部分组成。本系统开发了一个蒙古文显示输出系统，最终用户在 Windows 环境下蒙汉文同显窗口里进行编辑。

国家高技术智能计算机有关专家认为，这一系统的开发虽然刚刚起步，但有着广泛的应用价值，研究成果也表明，蒙古语语言学研究已达到世界先进水平。

（二）英-蒙机器翻译系统的研究

内蒙古大学计算机学院从 1998 年开始启动了国家自然科学基金项目“英-蒙机器翻译系统的研究”（项目编号为 69765001）。本系统是以英语为源语言，以蒙古语为目标语言的机器翻译系统，利用了基于规则的翻译方法。由于该系统是第一个英-蒙机器翻译系统，因此初步建立了一个中学初中英语范围内的英-蒙机器翻译系统 EMMT 模型。

本系统由机器词典和源语言分析与目标语言生成等三部分组成。机器词典包括单词词典和成语词典两类。乔姆斯基提出语言有“表层结构”和“深层结构”。后来国内一些学者提出语言还存在着一个“底层结构”。他们认为“表层结构”是语符结构，包括音符和字符。“深层结构”是语汇结构，包括词汇和句法。“底层结构”是语义结构，包括意义和情感。但是由于人力、物力的原因在此系统中没考虑“底层结构”。在翻译过程中使用到三种规则，分别是短语结构规则、短语结构规则转换和句法结构转换规则。

EMMT 系统的翻译过程中是：

（1）把源语言的表层结构分析成其深层结构，主要依靠源语言的短语结构规则和短语结构转换规则；

（2）把源语言的深层结构映像成对应的目标语言的深层结构，主要依靠源语言和目标语言间的句法结构转换规则；

（3）根据目标语言的深层结构生成其表层结构，主要依靠目标语言的语法规则。

后来此系统不断改进以基于规则的方法为主的同时，又以语料库的方法来补充和修正。

（三）基于实例的汉-蒙机器翻译系统

2003 年，内蒙古大学、中国科学院计算技术研究所和北京大学计算语言研究所联合申请了国家 863 计划课题“汉蒙机器翻译系统”（项目编号为 2003AA115510）。该项目根据以往的尝试，这次的汉蒙机器翻译系统采取了基于实例的翻译方法。考虑到以前研究的汉蒙双语对齐的技术基础，采用了基于对齐的 EBMT 系统。

系统的处理步骤如下：

1. 分词：将待翻译的句子切分成以词为单位的片段。在本系统中，汉语的分词采用了中科院技术计算所研发的 ICTCLAS 汉语分词系统。蒙古语按空格分词，不做特殊处理。

2. 对齐：将双语语料库中的汉蒙句对进行词对齐。这里是利用汉蒙双语词典及共现率为基础的方法进行对齐。对齐后，将双语语料库转换为实例库。

3. 实例搜索：从实例库中搜索所有最接近的实例。这一步的主要内容主要包含相似度的计算和搜索。

4. 片断匹配：在待译句子和实例中查匹配和不匹配片断。

5. 片断分割：根据匹配和不匹配片段确定翻译结果片断。

6. 片断组合：将翻译片段组合成翻译结果。

7. 评价：从候选的翻译结果中选择最佳的翻译结果。

对于不能从实例库中获得的翻译，可以利用双语词典进行翻译。也就是说，如果一个待翻译的句子无法找到接近的实例的话，将会退化成基于词典的翻译。双语词典包含源语言词、源语言词性、目标语言词、目标语言词性以及一个致信度。

在基于实例的汉蒙机器翻译方面，对齐语料库的规模和词对齐效果直接影响到翻译的质量。该课题的研究没有因课题的解题而结束，仍在继续扩充对齐语料库和探索更好提高词对齐效果的方法。

（四）基于模板的英-蒙机器翻译系统的研究

基于模板的英蒙机器翻译系统是一个以英语为原语言，以蒙古语为目标语言的；以基于范本的翻译方法为主的机器翻译实验系统，是内蒙古大学计算机学院的硕士研究生吉日木图于 2005 年的硕士学位论文中完成的，由敖其尔教授指导。

本系统不完全是基于模板的机器翻译系统，是在个别的地方试着用基于范本的方法解决。即：

1. 系统中对一些完整的简单句的翻译：用了基于规则的翻译方法。通过英语语法分析得到英语的语法树，然后按照英语和蒙古语之间的转换规则把英语语法树转换成蒙古语的语法树，再从蒙古语的语法树生成复合蒙古语语法的蒙古文句子。

2. 对常用的复合句或一些固定短语的翻译：用了基于模板的翻译方法。利用现有的模板进行模板匹配，得到比较准确的译文。初级阶段，人工生成模板，然后通过基于词典的机器学习之后从现成语料库中自动提取模板。

3. 对于模板和规则都不能正确处理的句子的翻译：本系统用基于 HMM 模型的蒙古文生成方法，采用词典驱动模型和 HMM 模型从单语料生成蒙古文译文。

从本系统的翻译结果来看，系统对简单句和并列复合句翻译译文的结构和语法正确率已经达到了一定水平。但是对于从属复合句、虚拟语气以及结构不完整句子的翻译效果还很差，原因在于本系统中所搜集的模版、语料库和双语词典的规模还很小，因此今后的主要工作是扩大模版、语料库等翻译资源的规模，提高系统的总体性能。

（五）基于统计的蒙-汉机器翻译系统

本系统是以蒙古语为原语言，以汉语为目标语言的基于统计的机器翻译实验系统，是内蒙古大学计算机学院的硕士研究生娜步青于 2006 年的硕士学位论文中完成的，由高光来教授指导。

该系统中收集了一定规模的语料库，搭建了蒙汉系统翻译模型、汉语语言模型与系统译码器，初步实现了一个基于统计的蒙汉机器翻译实验系统。语料库中包含中小学九年义务教育语文课本中的翻译浅显易懂的蒙汉对照文章、小部分新闻报导以及人工翻译的计算所机器翻译总库，共包含蒙汉对齐句子 9629 句对，蒙古文句子中蒙古语字数约 10、汉语句子中的汉语字数约 11，平均句长为 10 个单词。通过对蒙文复合词的切分、对翻译词典进行扩展等方法，提高了系统的性能。

此翻译系统在简单句翻译译文的结构和语义正确率相对高，这与蒙汉简单陈述句型相似、位置变换不大有一定的关系。但是对于从属复合句、虚拟语气以及结构不完整句子的翻译效果还很差。

虽然该实验系统在蒙古语语料预处理和蒙古语与汉语的翻译模型的建立方面做了有益的尝试。实验结果表明,与基于统计的蒙汉机器翻译相关的模型设计是合理的。该实验系统为今后的工作奠定了基础。但是该系统今后还要进一步进行研究,例如扩充双语语料库、增加翻译特征(如考虑位置变化特征等)和提高系统总性能等。

(六) 国外的蒙古语机器翻译情况

国外,主要是在日本进行日语和蒙古语之间的机器翻译研究。日语和蒙古语之间的机器翻译主要以新蒙古文和日文的机器翻译为主。日本的諏访东京理科大学(Tokyo University of Science, Suwa)等大学的江原晖将等人在进行关于蒙古语到日语的机器翻译研究。因为蒙古语和日语在语序上有相似性。蒙日机器翻译系统由源语言的形态分析、词汇转换和目标语言的生成组成。

由于蒙古语和日语的动词和否定形式的生成顺序不同,因此本系统针对此问题采用了传统的规则转换方法。

在目标语言生成部分,对译语的选择上本系统采用了统计的方法。

总体上蒙日机器翻译系统主要精力放在了选择译语的统计方法上。本系统以后的主要任务是扩充蒙日双语词典和双语平行语料库的扩充,从双语语料库获取更多的翻译知识,提高选择译语的性能和提高系统整体的翻译能力。

五. 总结

从以上的综述来看,蒙古语的机器翻译研究还处于初始阶段,还存在着很多问题,例如研究人员相对少,技术滞后,资源共享不够,经费缺少,规模小,到目前为止几乎没有产品化的成品,由于市场小,所以很难自食其力,得到经济效益。因此在很长一段时间内还是主要靠国家的投资来进行研究。并且只有经费是不行的,还需要各单位和个人的努力。必须注重培养人才,吸收前沿技术,与时俱进。相信有一天蒙古语的机器翻译研究在各方的努力下会有繁荣之日。

注释

①摘自冯志伟 2004 年版的专著《机器翻译研究》

②改编自王宜《机器翻译的回顾与展望》

参考文献

- [1] 冯志伟. 机器翻译研究 [M]. 北京: 中国对外翻译出版社, 2004. 1-2.
- [2] 刘群. 机器翻译技术的发展及其应用 [J]. 术语标准化与信息技术, 2002, (1). 27-30.
- [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] 詹卫东. 计算语言学基础研究内容与方法 [M]. 课件, 2006 年.
- [5] 王宜. 机器翻译的回顾与展望 [J]. 中国本地化网, 2004 年 9 月.
- [6] 俞士文. 计算语言学概论 [M]. 北京: 商务印刷馆. 2004 . 237-286.
- [7] 王·斯日古楞, 敖其尔. 英蒙机器翻译系统的设计 [J]. 内蒙古大学学报, 2003, (5). 582-587.
- [8] 珊丹, 敖其尔, 孟和吉雅, 斯琴高娃. 英-蒙机器翻译模型的探讨 [A]. 中国少时民族语言文字现代化文集 [C]. 1999. 333-338.

- [9] 吉日木图. 基于范本的英蒙机器翻译系统的研究 [M]. 研究生毕业论文. 2005.
- [10] 娜步青. 基于统计的蒙汉机器翻译系统 [M]. 硕士研究生毕业论文. 2006.
- [11] 娜步青. 基于统计的蒙汉机器翻译系统研究 [J]. 内蒙古农业大学学报, 2006, (2). 153-155.
- [12] 那顺乌日图, 刘群. 面向机器翻译的蒙古语生成 [A]. 自然语言理解与机器翻译 [C]. 北京: 清华大学出版社. 2001.
- [13] 那顺乌日图, 刘群, 巴德玛敖德斯尔. 关于“汉蒙机器辅助翻译系统”. ALTAI HAKPO(JOURNAL OF THE ALTAI SOCIETY OF KOREA) . 2001. 11.
- [14] 江原暉将, 早田清冷, 木村展幸. 茶釜を用いたモンゴル語から日本語への機械翻訳.

The Mongolian Machine Translation System

Murengaowa

(Academy of Mongolian studies, Inner Mongolia University, Huhhot 010021, China)

Abstract: In this paper, the writer has described the historical advances of machine translation technology and implementing methods, and gave a simple introduction of several typical machine translation systems of Mongolian.

Key words: Machine translation; Method of machine translation; Mongolian machine translation

收稿日期: 2007-5-16;

作者简介: 牧仁高娃 (1983-), 女, 蒙古族, 内蒙古赤峰市巴林右旗人。内蒙古大学蒙古学学院 2005 级硕士研究生, 主要从事蒙古文信息处理方面的学习与研究。