迈创英汉新闻机器翻译系统的研制

孙建军　鲁媛媛　梁　屹

北京迈创语通软件有限公司
multran@263.net

摘要：本文扼要地介绍了迈创英汉新闻机器翻译系统的实现过程，并从语言规则的角度对系统的实现技术，系统存在的问题以及解决问题的方法进行了简单地描述。同时我们也对系统在实际过程中的应用给予了说明。

Abstract: In this paper, we briefly introduced the development of Multran English-to-Chinese News Machine Translation System. In the view of the based-rule translation technology, we described the realization of the Multran technology, the problems existing in the Multran system and the approach to the problem. And we also indicated the application of the Multran system.

迈创英汉新闻机器翻译系统(以下简称为迈创新闻系统)是我公司从 2007 开始研制开发的. 该系统是在我们原有的迈创英汉口语机器翻译系统的基础上进行的,能够自动实现英语到汉语的转换. 目前，该系统已经能够对一般英语新闻文章进行翻译，译文有着良好的可读性。下面，我们就对迈创新闻系统的一般情况， 系统的技术特点，系统存在的问题和解决方法给予如下简要地介绍。

1. 迈创新闻系统简介

迈创新闻系统的特点就是对英语新闻信息的自动处理。该系统主要由系统实现程序，电子词典和迈创语言知识库等三个部分组成。在系统实现程序上，系统采用标准 C 语言编程，编译后的带调试信息的程序代码为 200K 左右，优化的嵌入式编译程序为 72K。 在系统数据规模上，系统的电子词典包含的英汉词语有 10 万余条，短语，片语以及词规则有近 60 万余条；系统语言知识库中的词法、句法语义规则有 1 千多条，可以标注的词法、句法语义信息有近千个。系统数据未压缩情况下有 24M 字节左右，压缩后可达到 8M 字节左右；在系统扩展方面，我们有一个专业的开发团队，虽然人数不多，但都是熟悉语言调试过程的英语语言专业技术人员，有着多年的丰富的调试经验。这在很大程度上，保证了系统的调试质量，为译文的可靠性提供保证。在系统应用方面，我们的系统是为数不多的能在市场上得到应用的产品。尤其是在一些便携式和嵌入式产品当中，我们的系统得到广泛应用，如在学习机、平板电脑、电子扫描翻译笔上都有应用,得到广大用户支持,也给我们带来了一定的经济效益，为我们的持续发展创造了条件。

2. 迈创新闻系统的技术特点

迈创新闻系统是采用语言规则方式来实现不同自然语言之间自动转换的。系统使用的语言翻译技术是我们自行研制的语言信息多功能逻辑运算分析技术。该技术采用数据驱动方式，语言知识数据和系统实现程序相互独立，通过借助于语言信息多功能处理函数实现对语言深层次信息处理。利用这种技术我们曾先后研制开发了英汉，汉英，日汉和俄汉等系统，并使用一个翻译引擎来实现多种自然语

言之间的自动转换，在改善系统语言翻译质量，提高系统实现效率，完善系统开发环境，增强系统移植能力，扩大系统应用等方面都表现出了良好的效果。

在系统语言翻译质量方面，迈创新闻系统通过提高系统对语言问题的解决能力以及结合对语言知识的有效处理使新闻系统的语言翻译质量有了一定的改善，同时也为系统的实际应用创造了条件。迈创新闻系统对自然语言的形式化描述是由专门的语言学家采用先进的语言信息深层次处理技术来实现的。这不仅使系统在语言问题上具有较强的解决能力，能够改善系统的翻译质量，同时也使得系统的数据规模得到了有效控制。

在系统实现效率方面，我们通过采用语言数据信息位码整合技术和数据信息位码压缩技术来加快系统的运行速度，减少系统占用空间。 系统的数据规模可以根据需要压缩在 7M 到 20M 字节之间，使系统能够在不同的平台环境下得到应用，适用于多个信息处理领域。

在系统的开发环境方面，迈创新闻系统的语言知识和运行程序是相互独立的。这就有可能使迈创系统在开放的语言学工程技术的基础上为语言学家建立了一整套的系统开发工具，从而可以使专门的语言学家来实现语言知识库的建立和不断地扩充，可以有效地降低系统实现的复杂难度，为系统的可持续发展提供保证。

在增强系统移植能力的方面，迈创新闻系统的机器翻译引擎是采用标准 C 语言编程，能够在多种操作系统平台下运行，具有兼容性好，移植能力强和翻译速度快等特点。

3. 迈创新闻系统现在存在的问题及解决方法

我们知道，自然语言的信息处理涉及到多个方面，尤其是其与人类思维的紧密联系，使得我们在对自然语言信息的自动化处理方面临许多难题。现有的机器翻译系统研发技术大部分都还局限在基于句子结构分析范围内的语言信息自动化处理。而对篇章分析，背景知识运用等还处于研究阶段，实际应用的技术不多。局限于语句范围的语言信息处理也会因使用的语言信息处理技术不同而存在这样或那样的问题，而这也直接关系到译文的翻译质量。

我们从事机器翻译系统的研发已有多年，在系统研发中所涉及到的专有名词处理，兼类词处理和语言消歧处理等方面仍有许多工作需要去做。这一点在新闻翻译系统的开发过程中显得尤为突出。首先，专有名词在新闻报导中出现的频率是比较高的，而且种类繁杂。从人名，地名到以各种词汇组合形式出现的机构组织名称，每一种情况处理不到位，都会给译文的翻译质量带来影响。当然，最好的处理办法是把它们都收录到电子词典中去。但这在现实中是不可能的。专有名词的特点就是不断地变化，不断地推陈出新。虽然从技术实现的角度来看，专有名词的识别是不难的，可以根据单词的大小写来进行识别。但当识别出以后如何给出正确的语义信息和如何对多个词汇的组合给出合适的命名是解决专有名词问题的关键。因为它们直接关系到其它词汇的译文选取和句子的可读性。在这方面，我们还有很多工作去做。

其次，新闻报道中经常使用的常用词汇，也导致兼类词的出现频率也会比较多。有些兼类词判别容易，有些判别的语境比较复杂，一旦判别错误就会导致句子的翻译结果出现偏差。针对这一现象，我们采取的策略是，容易判别的尽量去判别，复杂判别不出来的就保留其兼类词的成份，然后提供给系统去针对每种词类去分析，最终选择一个合适的结果。目前这种技术在实际应用中取得了良好的效果，减少了因兼类词对语言翻译所造成的影响。

最后，自然语言的消歧问题是我们开发机器翻译系统中遇到的最大问题。可以说语言消歧贯穿了机器翻译系统研发的

整个过程。从词汇歧义，词组歧义到短语结构歧义和句子结构歧义，语言歧义问题是无处不在，成为影响翻译质量的一个重要因素。尽管可以通过上下文信息解决一部分歧义问题，但歧义问题所涉及的一些复杂信息，如篇章知识，文化背景知识等还无法处理，这就给歧义问题解决带来了很多困难。就目前对自然语言的认识以及使用的翻译技术来说，全面解决语言歧义问题还有待时日。因此如何去发现这些问题，然后去思考这些问题，进而找到解决问题的方法是我们目前能做的一种明智选择。这同时也为以后问题的解决提供了有效的素材。从我们目前的研发工作来看，我们认为找到语言歧义和如何包容语言歧义，提高译文的可读性要比完全解决语言歧义更为现实，也就更容易贴近用户。如我们的系统在对词语歧义问题处理上，能够利用词法、句法、语义信息进行消歧的，系统会尽量做出选择。有些多义词经过分析处理后，仍然有多个译文选项存在，系统会自动给出 2-3 个译文。这虽然在一定程度上会影响译文输出结果的流利程度，但在帮助用户理解原文上有很大益处，在实际应用中受到普遍欢迎。此外，我们也计划针对语句的短语结构歧义和句子结构歧义采取类似的处理。对于一些目前无法解决的语句结构歧义问题，可以为原文提供 2 到 3 个翻译结果，从而帮助用户进一步理解原文。这一方面在消岐知识不健全的情况下,可以帮助用户解决一些问题；另一方面也为发现问题和最终解决问题创造条件。

4.总结

我们长期以来一直是从事基于语言规则方式的自然语言处理技术研究。在目前基于语料统计技术的机器翻译系统研制已经成为一种主流的情况下，语言规则系统开发究竟能走多远一直是人们比较关心的话题。甚至有人认为语言规则系统已经没有了出路。当然这也和过去语言规则系统开发过程中出现的一些误区有关。在我们看来，语料统计和语言规则是自然语言应用表现的两个不同方面,二者相辅相成,共同代表着自然语言“约定俗成”的特性。语料统计通过借助于人类的翻译成果，在译文流利度方面显现出一定的优势。而语言规则是在模拟人类语言思维机制，是对人类智能的巨大挑战，在忠实地反映原文方面体现出更加实用的效果。相信两者会在今后的自然语言信息处理中得到共同的发展。

参考文献：

[1] 刘开英, 郭炳炎. 自然语言处理[M]. 科学出版社，1991.
[2] 冯志伟. 数理语言学[M]. 知识出版社，1985.