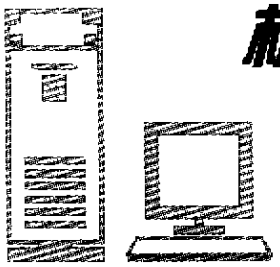

机器翻译技术的发展及其应用

◆ 刘　群(北京大学计算语言学研究所,中国科学院计算技术研究所)

关键词:机器翻译;中文信息处理;计算语言学;自然语言处理

摘　要:本文结合一些典型的机器翻译系统,介绍了近年来机器翻译技术的进展情况,探讨了机器翻译技术的发展趋势,最后介绍了目前机器翻译的实际应用情况。

Advances of Machine Translation Technology and its Applications

LIU Qun

Keywords: machine translation, Chinese information processing, computational linguistics, natural language processing

Abstract: By decrypting several famous machine translation systems, the paper introduces the recent advances of machine translation technologies, and discusses the trend of the future development. The application of machine translation is introduced at last.

一、引　言

几乎从计算机诞生之日起,人们就试图利用计算机来进行自然语言的翻译工作。不过,机器翻译的先驱者们也许没有想到,人类进入二十一世纪后,计算机早已渗透到人类生活的方方面面,计算机技术的发展已经远远超过了他们所能够想象到的程度,而机器翻译却依然没有达到理想的水平。

纵观机器翻译的发展历史,应用一直是推动机器翻译发展最主要的动力,尽管机器翻译经历了如此曲折的发展历程,尽管机器翻译的质量到现在为止还远不能令人满意,机器翻译技术的发展却一刻也没有停止。尤其是近年来,Internet 的普遍应用,世界经济一体化进程的加速,使得人们对于机器翻译的需求也空前增长,机器翻译的研究也迎来了一个新的发展机遇。

本文从技术与应用两个方面对机器翻译技术近年来的发展及其应用情况作一个回顾。

二、机器翻译技术的进展

与其他自然语言处理技术一样,机器翻译技术从总体上也可以分为基于规则和基于语料库两大类。关于这两类方法的优点和缺点,很多文章都作过详细的分析,这里不再重复。本文拟采用个例分析的方法,通过介绍几个典型的机器翻译系统,向读者展示机器翻译系统的一些新的技术和新的思路。希望读者能够了解到机器翻译技术的最新进展并从中获得启迪。

1. IBM 的统计机器翻译模型

IBM 公司开发的一个英法机器翻译系统采用的基于统计的机器翻译方法。其基本思路是，把机器翻译看成是一个信息传输的过程，用一种信道模型对机器翻译进行解释。系统需要翻译的是一段源语言文本 S，该模型假设 S 是由一段目标语言 T 经过某种形式的编码得到的，将 S 翻译成 T 的过程就是一个解码的过程。根据 Bayes 公式可推导得到：

$$T = \max_{T} P(T)P(S \mid T)。$$

这里，$P(T)$ 目标语言的文本 T 出现的概率，称为语言模型。$P(S \mid T)$ 是由目标语言文本 T 翻译成源语言文本 S 的概率，称为翻译模型。语言模型只与目标语言相关，与源语言无关，翻译模型与目标语言和源语言都有关系。对这两种模型的不同推导，可以得到不同形式的基于统计的机器翻译数学模型。而建立这些模型的所有数据，都必须来源于对大规模语料库的统计。

IBM 公司 Peter Brown 等研究者按照这种方法，以英法双语对照加拿大议会辩论记录为双语语料库，开发了一个英法机器翻译系统。从他们发表的文章来看，实验的结果相当不错，已经超出了传统的基于规则的翻译系统。不过，这种成功可能与两方面的因素有关：一是英法语言相当接近，语序的调整很小；其次有很好的对齐语料库作为基础。可惜的是，这项工作后来并没有继续下去，其他类似的工作也没有再重复过这么好的结果。

2. Microsoft 公司的多国语机器翻译项目

微软研究院（美国）开展了一个多国语机器翻译项目，该项目采用的是一种基于规则的方法。

在该项目中，每一种语言有一个独立的开发小组，开发该种语言的分析器。系统提供统一的分析器开发工具平台——G 语言。G 语言是一种专门为自然语言处理设计的形式语言，采用 C 语言的语法形式和类似 Lisp 表的数据结构。G 语言代码可以快速转化成 C 语言代码，经过编译后就直接得到了可执行程序的句子分析器。开发句子分析器的过程就是编写 G 语言代码的过程，所有的句子分析规则都以 C 语言代码的形式体现出来。句子分析的结果是一种逻辑表达式。各种语言的逻辑表达式并不统一，但采用相同的语法形式。

在该系统中，并没有专门的转换规则，转换规则的获取是利用双语语料库中全自动进行的。目前他们采用的语料库是微软的产品说明书，是句子对齐的，数量相当庞大。对齐的方法是先采用他们自己开发单语的分析器对两种语言的句子进行分析，再对得到的两个逻辑表达式进行对齐，就可以得到大小不等的各种对齐短语块，而这些对齐的短语块可直接用于语言的转换。

在该系统中，虽然也利用了双语句子对齐的语料库，但是并不对这种语料库进行任何人工标记或校对，从语料库中获取的翻译知识完全是动态的，一旦某种语言的分析器进行了修改，这种翻译知识就需要重新生成。

目前该系统经过多年的开发，已经达到了相当的规模，虽然还没有作为正式产品推出，但其实力不容低估。

3. AT&T 公司的语音翻译系统

AT&T 公司开发的语音翻译系统由语音识别、机器翻译、语音合成三部分组成。语音识别与语音合成与本文无关，这里不作讨论。它们在机器翻译部分采用的算法非常独特，这里作一个简单的介绍。

该系统所采用的核心的数学工具叫做“中心词转录机”（Head Transducer，以下简称 HT）。这实际上是有限自动机（Definite State Machine，以下简称 DSM）的一种扩充，与有限状态自动机的区别主要体现在两个方面：一是 DSM 的输入是从左到右进行的，HT 的输入是从中心词向左右两个方向扩展的；二是 DSM 只

有输入没有输出，只能识别不能翻译，而TM在输入的同时产生输出，可以用来翻译。从表达能力看，单个HT的表达能力比单个DSM并没有提高。众所周知，单个DSM的能力等价于正规语法，但在这个系统中，翻译知识是用一组HT来表示的，HT之间可以嵌套（一个HT识别的结果可以作为另一个HT的一条边），因此其表达能力弱等价于上下文无关语法，比正规语法更强。

该系统的所有翻译知识（包括词典、分析规则和生成规则）都用一组带概率的HT来表示。这些知识完全从语料库中自动获取，获取的过程无需人工干预，但获取的结果（就是HT）很直观，可以由人进行调整。HT的表示是完全基于词的，不采用任何词法、句法或语义标记。

整个知识获取的过程实际上就是一个双语语料库结构对齐的过程。句子的结构用依存树表示（但依存关系不作任何标记）。他们经过一番公式推导，把一个完整的双语语料库的分析树构造并对齐的过程转化成了一个数学问题的求解过程。这个过程可用一个算法高效实现。得到对齐的依存树后，很容易就训练出一组带概率的HT，也就得到了一个机器翻译系统。

这种方法比较适合于语音翻译这种领域比较受限、词汇集较小的场合，对于大规模的文本翻译并不合适。但这种做法对我们开拓思路还是非常有借鉴意义的。

4. CMU的Pangloss系统

PANGLOSS系统是美国卡内基梅隆大学研制的一个西班牙—英语的机器翻译系统。该系统采用一种多引擎的策略。

该系统采用三个翻译引擎。一个是基于知识（规则）的翻译引擎，一个是基于实例的翻译引擎，一个是基于词语转换的翻译引擎。

其中基于实例的机器翻译引擎不需要任何语言结构知识，需要的语言资源有一个双语例句库、一部双语电子词典和一个目标语言的同义词分类器。其中的双语例句库的规模达到了72万西班牙—英语语句对，主要来自于UN Multilingual Corpus, Linguistic Data Consortium；而目标语言的同义词分类器是从WordNet中提取得到的，连同双语词典一起来寻找源语言和目标语言语句中词语的关联，根据统计，该实例引擎对于不限制领域的输入能够达到70.2%的覆盖率。

基于实例的翻译引擎实现方法：

1. 通过查找例句库的索引寻找同输入匹配的最长组块（chunk）；

2. 进行语句片段的对齐，确定组块的译文。

对于一个输入语句，每一个翻译引擎对其任何片段都可以产生译文，然后把这些产生的译文放到一个统一的线图（Chart）中，最后根据一个统计模型来决定线图中的最佳路径作为译文输出。这样做可以尽量结合各个翻译引擎的优点，有利于产生最好质量的译文。

通过上面介绍的四个系统可以看出，这些系统所采用的技术可以说是各有特色，丰富多彩，每一种方法都有其成功之处，但也都有自身的弱点。从发展趋势看，规则和统计相结合的方法应该是机器翻译技术的一个发展趋势。笔者认为，这种结合可以体现在三个层面上：第一是知识表示层面，要求知识表示形式既能够表达深层的语言学知识，又能够与某种统计模型相结合，具有很好的鲁棒性；第二是知识获取层面，知识库应该可以通过大规模语料库自动获取，同时，这种知识又是直观的，可以为语言学家所理解和修改；第三是知识运用层面，这主要是指的多引擎方法，多种不同的翻译引擎互相补充，互相取长补短，总体上可以达到更好的翻译效果。

三、机器翻译技术的应用

从机器翻译的水平看，虽然目前还远未能达到全自动高质量的理想目标，但目前机器翻译还是可以满足人们多种多样的翻译需求，在

各种领域获得了广泛的利用。

从应用类型看，机器翻译可以分为四种类型：

1. 信息发布型

这类系统主要为信息发布者提供翻译服务，主要的翻译内容是新闻、法律、公告、产品说明书等等。这类用户需要准确地将自己的意思用另外一种语言表达出来，因而要求很高的准确率。为了达到很高的准确率，一般有两种形式。一是采用受限语言。在严格受限的领域内，机器翻译可以达到很高的翻译质量。典型的如加拿大 TAUM－METEO 天气预报翻译系统。某些公司的产品说明书也是要求采用一种受限语言，这样不仅可以保证写出的文字非常浅显易懂，同时也使得机器翻译变得比较容易。另一种形式是采用机助人译的方法。如采用翻译记忆(TM)技术的计算机辅助翻译系统，又称翻译工作站(Translation Workbench)，目前这类产品已经相当成熟，除了提供一般的翻译记忆功能以外，还可提供文件格式的分解与合成、术语库管理、翻译项目管理、语料库加工与对齐等一系列辅助翻译工具。目前此类产品已经形成了比较大的产业规模，得到了专业翻译人员的普遍认可。这方面最著名的系统是 Trados 系统，中文译名是“塔多思”，号称其翻译解决方案销量超过 40,000 多企业用户，占整个翻译软件市场的 70%。

2. 信息吸收型

这类系统主要是为那些不需要了解准确的含义，只需要浏览其大意的用户提供的。这一类系统随着 Internet 的推广而得到了迅速的发展。在这类系统的帮助下，一个完全不懂外语的人，也可以大致看懂外语网页的内容，这对于很多用户有相当大的吸引力。比较著名的此类系统如国内的“看世界”网站和国外的 WordLingo 网站等。

3. 信息交流型

这一类系统需要为那些进行一对一的交流的人们提供翻译服务。这类系统又包括口语翻译系统和文字翻译系统。口语翻译系统目前已经取得了一定的进展，但由于语音技术的限制，目前一般都只能限于小词汇集和非常受限的领域，如旅馆预定、机票查询等。文字翻译系统没有口语识别的问题，可以用于更广泛一些的领域，如旅游翻译、电子邮件的翻译系统、在线聊天(Chat)翻译等等。这类系统的主要特点是：翻译的内容主要是口语而不是书面语，表达比较随意；大多面向特定领域；翻译的实时性要求更高，人机交互更为复杂等。

4. 信息存取型

这一类系统指用于多语言环境下信息检索、信息提取、文本摘要、数据库操作等目的的嵌入式机器翻译系统。由于目前 Internet 的迅速发展，这类系统发展也很快。例如跨语言的信息检索目前已成为信息检索领域的一个重要研究课题。

以上这些应用形式中，有些已经相当成熟(如计算机辅助翻译系统)，有些却还是处于实验阶段。不过，随着机器翻译技术的发展和机器翻译水平的提高，越来越多的翻译系统将走向实用，一些新的应用形式也将被创造出来。

四、结束语

如果要将各个历史时期人们对机器翻译的期望值进行比较的话，那么，除了早期人们的盲目乐观情绪以外，现在应该是人们对机器翻译最为看好的时期，而且这种看好是建立在对机器翻译的能力有一个客观认识的基础上的。笔者对于机器翻译持一种非常乐观的态度。笔者认为，机器翻译技术现在正处在一个从量变到质变的积累时期，而这个质变也已为时不远。也许不久以后，机器翻译软件就会像现在的文字处理软件和多媒体播放软件一样，成为大家日常生活中一种常用的工具，而语言的障碍对我们日常交流造成的困扰也将在很大程度上被克服。