英汉机器翻译系统的建造
—用于英语词典翻译出版的专用系统

郑保山 刘 群 张 祥

中国科学院计算技术研究所 北京 100080

摘要 本文从人工翻译和机器翻译的经验出发，引入数据仓库和数据挖掘技术建造语料库，提出一个采用模板技术的译语精确生成和机助人译结合的动态机器翻译系统，专门用于英语词典的翻译出版，促使机器翻译走向实用化，初步研究取得了较好效果。

关键词 机器翻译 数据仓库 数据挖掘 译文模板 机助人译 英语词典

Construction of an English-Chinese MT System for Translation of an English Dictionary to Publish

Zheng Baoshan Liu Qun Zhang Xiang

Institute of Computing Technology, CAS Beijing 100080

Email: bszheng @jlonline.com

Abstract From the experience of artificial and machine translation, this paper introduces the data warehouse and data mining technology to construct corpora, suggests a machine translation system that the accurate generation of text to be translated using a template technology combines with person translation under machine help, useful in the translation and publication of English Dictionary into Chinese one. Results of the preliminary research have gotten better effect.

Keywords Machine Translation Data Warehouse Knowledge Discovery in Database Translation Template Person Translation under Machine Help English Dictionary

一、问题的提出

应出版社的要求，翻译一本叫做《Computing Dictionary》的计算机方面的词典。词典的特点是：

(1) 电子版为带标文本文件 comp.txt，容量为 1.4 Mb，1 万多个词条，包括释义、例句、注释、附注和引文等各种标记 20 种，共 30595 项。

(2) 词条项的表示具有比较复杂的形式，含有词的用(规范词)、代(代用词)、属(上位词)、分(同义词、近义词和反义词)和参(参照词)等，使用了一些符号如"＝""＊"代替说明。

(3)句子成分不全，多为无主句或为短语，是在词条下省略词条而表述的句子。

(4)每个词条下的内容都不一样,有简有繁,而且每项标记的内容并不就是一个句子或短语,是若干项组合的结果,往往变成句子需要某些项的合并。

(5)电子文件既然是一个出版物,就允许一定的差错率,无论是内容还是标记符号都有疏忽或来不及修改的地方,因而数据的清洁仍然是一个任务。

(6)翻译完了之后要求形成带标文件的形式,既能提取出词条及其相应的汉译内容,个别项还需要提取出汉英对照的形式。

对于这样的一种源语形式,能不能用机器翻译来完成?搞机器翻译的人,不用机器翻译而用人工翻译,这也太具有讽刺意味了。在这样的背景下,为了探索机器翻译走向实用的方法,我们从这个特例开始,避开源语的变化,专门研究源语文本确定的实用机器翻译系统。

二、机器翻译系统的设计

对于机器翻译,文献[1]提出了从单句处理走向句群处理、建立新的知识系统和改进译文生成的一些可能引起技术变革的关键问题。文献[2]提出在自然语言受限处理领域的语言信息交流模板的新技术,要求既能及时地反映自然语言的发展变化,又能合理地限制自然语言表达的随机性。文献[3]又提出了一个全息全选全程翻译系统。文献[4][6]提出了重视人工翻译经验在机器翻译中的应用。文献[5]提出了将经验主义方法与传统的基于规则的理性主义的体系结合起来的混合策略。

看来,基于规则与基于语料库的结合、对源语加以合理的限制、加强语义网络的研究和译文生成采用模板等项技术,是当前机器翻译界正在探索和研究的内容。系统的设计思想如下:

(1)采用基于语料库和基于规则相结合的策略,编制知识获取的常用工具,比较全面地进行数据挖掘的研究。

(2)不拘于一种形式翻译。对于词条、词类、参照项和引文出处等,在建造语料库的同时在程序内译好;对于常用谓词和高频句,采用句子模板的方法翻译;对于不规范或覆盖不了的短语或句子,采用机助人译的方法翻译,再加上机器学习的机制,使系统随着用户的使用而提高。

(3)从词频统计结果,通过常规的机器翻译系统词典建立领域专用的专业电子词典,通过词组(句片)的统计,建立词组(句片)库,提高翻译的针对性。

(4)自动生成源语模板,仔细制作译文模板,力求译文的精确生成,从总体上降低用户机器翻译后的人工编改工作量。

(5)对于不能确切机器翻译的句子,提供机助人译平台,实现人机互补,就地人工翻译,并且建立机器翻译机制,一个句子一次人工翻译永久受益。

三、《计算词典》语料库的建立

从1997年起,数据仓库的方法得到越来越广泛的认同,在数据爆炸环境下找到了一种不为数据的汪洋所淹没的知识获取的方法。对于机器翻译,数据仓库(Data Warehouse)与语料库(Corpora)之间在本质上没有什么区别。数据仓库面向主题整合,随时间变化收集数据,是一种支持管理决策的数据结构形式[8]。它是先用数据库将数据形式化并组织起来,然后从中发现知识,因而提高了知识获取的起点[9]。数据仓库概念的内涵比语料库的内涵要大,更抽象一些。在归属上,语料库是数据仓库在语言学和计算语言学研究中的应用。在数据仓库中,注重数据的清洁度,数据垃圾太多,不会有正确的决策结果的。它还讲究对决策结果的专家评估,

注意发挥人的作用；讲究一个过程的反复处理，从不断修正中求得正确的结果。

本系统的源语有一个现成的半结构化带标文本文件，语料库的建立相对比较容易。可是，如上所述，数据的清洁度仍然有一些问题，需要严格校对、修改，特别是标记方面的错误，具有全局性影响，为此我们曾清理过二次。最后以基本词条、词组、缩写、同义词条、一词多义、兼类作为关键词，建立了 11263 项的 3.75 Mb 语料库，并分别提取出词条项 10557 组、释义项 10366 组、例句项 1461 组、注释项 217 组、附注项 157 组和引文项 383 组的单独文本文件，用于机器翻译。同时对各提取项进行编码，用于机器翻译后译文的装入合成。

《计算词典》语料库还包括数据挖掘生成的词频库、专业词典库、中心词（谓词）库、词组库、句片库、句子库和源语和译语的模板库等，总容量达到 200 Mb 以上。

四、语料库的数据挖掘

数据挖掘也称之为数据库中的知识发现（Knowledge Discovery in Database，KDD），是从大量数据中提取出可信、新颖、有效而又能为人理解模式的高级处理过程[8][11]。文献[7][10]已经把这一方法应用到机器翻译的研究中。数据挖掘在商业上是建立在数据仓库的基础上进行的，目的是支持商业老总们的决策。现在，这种方法很盛行，并且取得了可观的经济效益。世界各大数据库公司推出的数据仓库及其数据挖掘技术，关键是它的数据挖掘工具。

我们的数据挖掘，是以语料库及其提取出来的文本文件作为数据源，全面掌握源语的语言学和计算语言学方面的知识，以支持我们来建造机器翻译系统。数据仓库的建立方法和数据挖掘工具的商品化，值得我们借鉴。在数据挖掘之前，我们首先开发了很多用于语料库的数据挖掘工具。这个系统所使用的挖掘工具，主要是词频统计、句片边界的确定及提取、句型统计及句子模板的自动生成和中心词例句的自动装入等。具体挖掘流程如图 1 所示。

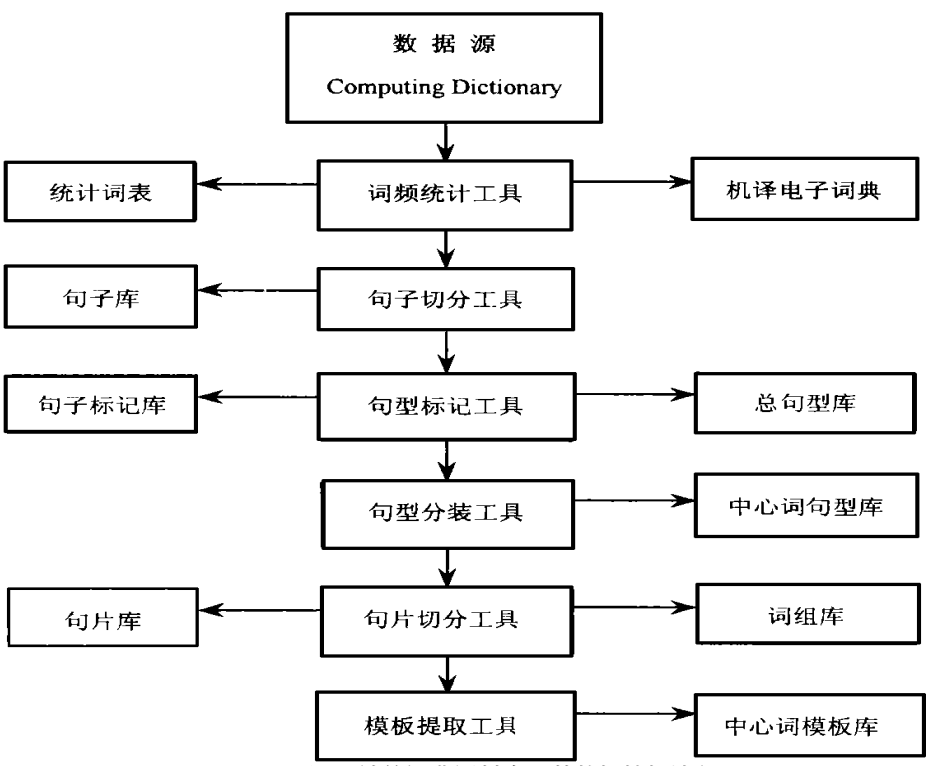

图 1　计算词典语料库及其数据挖掘流程

五、电子词典的建造

由于上述系统的设计特点，需要建造的电子词典包括基本词典、专业词典、句片库和词组库。在数据结构上，基本上都包括英文项、汉译项、词类参数、语义参数、统计频率和使用频率。

我们原来编有基本词典和计算机专业词典，但是在这里只作为知识库使用，词频和词组统计完毕后，用它自动生成系统的基本词典和专业词典。基本词典只包括动词、助动词、连词和介词，专业词典只包括由语料库统计出来的专业词汇。这样一来，系统的基本词典和专业词典规模都很小，基本词典收词 1779 个词，专业词典收词 7124 个。这样做的目的是与系统设计相配合，提高词典的专指性，做得好一些，实用一些。

词组库的建造基于句片的提取，句片中可构成词组的部分一定要形成词组，这关系到减少译文中的垃圾，消除歧义和译文的精确生成。例如：output-to-input signal strength ratio，输出对输入信号的强度比。从句片库中提取多字构成的句片，其汉义不能由其各自的汉义叠加而成者，做成词组。句片库自动收句片多达 14192 个，其中大部分不符合做成词组的条件，做成词组者仅为 3829 个，但对精确生成至关重要。

六、句型统计和模板的生成

释义部分有 11110 个句子，统计的句型为 8835 个。例句部分有 1807 句，句型为 1482 个。注释部分有 341 句，句型为 313 个。附注部分有 91 句，句型为 55 个。引文部分有 428 句，句型为 415 个。整本词典共有 13777 句，句型为 11100 个，用词量为 165874 个词。句型的重用率为 19.5%。统计词数为 8350 个，词的重复使用率平均每词为 19.4 次。

模板的制作是根据句型统计进行的。译文模板的制作原则是：

(1) 使用二次以上的句型都制成模板。

(2) 常用动词的各种成分分布的句型也都制成模板。

(3) 常用短语或词组，包括标题，制成模板。

模板的加工过程如下：

1. 英文模板是自动生成的，如

(screen) where the text and graphics are shown as black on a white background to imitate a printed page.

2. 由源语生成模板

([1])-where[2] and[3] are-shown-as[4] on[5] to-imitate[6] printed[7] .

3. 源语的操作模板为

([1, 1, 2])-where[2, 2, 5] and[3, 1, 8] are-shown-as[4, 1, 12] on[5, 3, 14] to-imitate[6, 1, 19] printed[7, 1, 21] 。

4. 译语模板人工制作为

([1])上把[2] 和[3] 显示成[5] 上的[4] ，用以模仿打印的[7] 。

5. 根据模板机器翻译结果为

(屏幕)上把文本和图形显示成白背景上的黑色，用以模仿打印的页。

6. 基于规则的机器翻译系统的译文为

(屏幕) 文本和图形在一个白的背景之后作为黑色的在哪里显示(以)模仿一个打印的页面。

七、机器翻译系统的构成

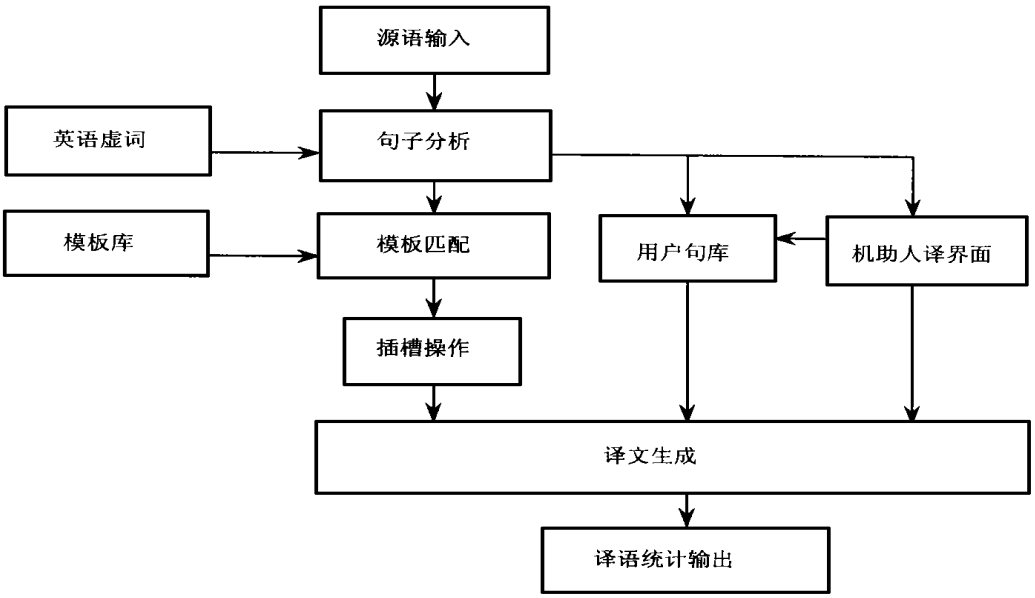

图 2　机器翻译系统框图

八、初步结论

1. 机器翻译源语的不确定性，为机器翻译的开发带来很大的难度。我们在语料库的建造方面借鉴了数据仓库及其数据挖掘的长处。现在每一个学者或专家的研究领域，已经从面上逐渐地撤退到一个点上，专家拥有的知识已经不适应于面上的需求了。对于涉及这样广泛领域的机器翻译来说，专家获取知识的方法需要与数据挖掘结合起来。我们在机器翻译技术的开发上，把重点转移到源语的处理上来，把不确定的东西搞清楚，在这样的基础上建立机器翻译系统。

2. 现在机器翻译存在的问题，除了上面所说的文不对题的现象之外，主要是词序不对和译文中不必要的垃圾太多。想用机器翻译代替人工翻译干点活，可是，活干完了一点没有省事，还是不如自己译来得好。这种结局往往使用户放弃了机器翻译的使用。我们采用了译文的模板技术(有点类似于格框架[12][13])，先把词序调整好，规则做在模板上，达到了译文的大部分比较精确生成的目的。有一句俗语说得好：之乎者也矣焉哉，安排好了当秀才。实词是随机开放的，虚词是可以穷举的。我们以谓语为中心词，以助词、连词和介词作为框架结构，让实词在这个框架结构中发展，避免了它的泛滥。框架中的实词做成空白，好填各种各样的实词，称之为实词槽（slots）。槽中有实词的槽号、地址指针、大小和语义参数。这样一来，就把机器翻译的主要操作简化为插槽的可控操作。

3. 模板不能解决机器翻译的所有问题，只能解决一部分问题。找不到相应模板的句子，退而做例句翻译操作，因为系统含有一个累积例句库，例句的数量远远大于模板的数量。上述二项都匹配不上者，系统自动降为机助人译平台，让用户参与。系统使用传统的机器翻译方法提供初步翻译的结果，并示出参考例句和参考译文模板。机助人译的结果，立即为机器所存储，作为系统的知识库之一，可以为系统所调用。所以，这样的机器翻译系统是一个由机器翻

译和机助人译过渡到总体上是机器翻译的动态系统。翻译的源语越多,用户参与的越多,机助人译的出现也就越少,机器翻译的性能也就越好。不管系统处于动态的何种程度,翻译的句子都是可用的,都能解决人工翻译的一部分负担。相信喜欢机助人译或写作平台的人,会更加喜欢这样的系统。

4. 由于在机器翻译系统的建造上侧重于领域内的语料研究,所以系统是面向领域的,具有定制相应机器翻译系统的能力。前提是系统先要有一个基本覆盖本领域的数据库,有现成的最好,也可以通过网上下载或录入扫描等办法来建立。当然,每次翻译源语之前都可以进行这样的语料统计和处理操作,去深入了解要翻译的内容,为系统积累知识。由于计算机软硬件的迅速发展,机器翻译的时间已经可以略而不计,因而,即使加入了这样的语料处理步骤,也还是切切实实地节省了人力。对于较大的翻译任务,如 1 万字以上,以先做好语料的准备为好。

5. 这个实验系统是专为翻译一本词典而设计的。现在出版社最怕用机器翻译,因为词典的结构比较复杂,需要翻译的内容与不需要翻译的内容混在一起,搞乱了更加不好收拾。为此,专门设计了可译内容与不可译内容的分开和合并程序,以及带标文件的输入和输出程序,进行了语料加工和复杂半结构化带标文本文件机器翻译的成套设计。

6. 本机器翻译系统的框架已经初步形成,训练效果比较理想。目前的主要工作是译文模板的制作,它需要人工精心翻译句子的配合,工作量比较大。模板积累得越多,系统也就越成熟。这项研究工作开始不久,还有很多问题需要探索、试验和完善,渴望得到各位的批评指正。

参 考 文 献

[1] 董振东. 机器翻译研究进展. 计算机世界, 技术专题 D2, 1998-4-13

[2] 刘莎. 21 世纪面向计算机技术的自然语言的变革. 计算机世界, 专题综述 D31, 1997-12-22 [3] 刘莎. 全息全选全程翻译系统. 计算机世界, 技术广角 D14, 1998-5-11

[4] 郑保山. 机器翻译技术及应用. 计算机世界, 技术专题 D4, 1998-4-13

[5] 王海峰等. 汉英双向机器翻译系统 BT863 的研究与实现. 情报学报, 1997, 16(5): 360~369

[6] 郑保山. 谈谈二次文献的机器翻译. 中国科技期刊研究, 1998, 9(3): 169~171

[7] 郑保山, 龚小芬.《精细石油化工文摘》数据库的自动建立. 计算机科学, 1998, (10), 第十五届全国数据库学术会议论文集, 102~103

[8] 朱廷劭. 数据挖掘——极具发展前景的新领域. 计算机世界, 产品与技术 C14, 1999-01-04

[9] 美国 SAS 研究所. SAS 数据仓库与数据挖掘——从业务数据中提炼决策支持信息的解决方案. SAS 系统介绍, 第十五届全国数据库学术会议文件, 1998

[10] 范建华, 李德毅等. KDD 和基于语料库的机译系统. 见: 计算语言学研究与应用. 北京: 清华大学出版社, 1995, 302~307

[11] 李德毅. 数据开采研究现状. 计算机世界, 专题报道 D5, 1998-03-02

[12] 刘开瑛, 郭炳炎. 自然语言处理. 北京: 科学出版社, 1991

[13] 冯志伟. 自然语言机器翻译新论. 北京: 语文出版社, 1995