DOI:10.13916/j.cnki.issn1671-511x.2002.03.034
2002年 5月 东南大学学报(哲学社会科学版) May. 2002
第 4卷第 3期 Journal of Southeast University (Philosophy and Social Science) Vol. 4 No. 3

机器翻译与计算机辅助翻译研究与探索

李 鲁

(东南大学 外语系,江苏 南京 210096)

[关键词]机器翻译;计算机辅助翻译;语法模式;面向数据处理

[摘 要]机器翻译和计算机辅助翻译在 20世纪已经取得了一些成就,但国内目前的一些翻译软件的机译质量仍未尽如人意,在翻译系统语法库和对于某些词项的词义定义方面还需进一步推敲,其中一个有效的方法就是添加或修改某些广义语法模式定义或指令。

[中图分类号]H31 [文献标识码]A [文章编号]1671-511X(2002)03-0175-05

一、机器翻译与计算机辅助翻译发展简史及其理论基础

机器翻译的设想最早是由法国科学家 G° B° 阿尔楚尼在 20世纪 30年代初提出。随后,在 1933年,苏联科学家 P° P° 特罗扬斯基提出借助机器进行翻译的详细步骤,并设计出由一条履带和一块台板组成的依靠机械原理进行翻译工作的样机;但他最终未能按样机制成具有实用功能的翻译机。

1946年,英国和美国的两位工程师 A° D° 布思和 W° 韦弗首次提出利用计算机来进行翻译,他们并于 1949年出版了《翻译备忘录》一书。此后的几年间,美国、苏联、英国等国不断有学者开始参与或关注机器翻译方面的研究。

1954年,美国乔治敦大学和国际商用机器公司(IBM)首次联合试验使用电脑机译系统,并将由 250个词组成的简单的俄文材料译成了基本上可以接受的英文。这次试验的成功标志着机器翻译系统的真正诞生。此后,美国、苏联、日本、意大利、比利时、英国、德国等国便掀起了机器翻译热。

1956年我国开始研究机器翻译,但由于随后出现的一些政治运动,尤其是几年后发起的文革运动,该项研究被长期搁置。直到 80年代初期我国的机器翻译研究才得以继续,并受到高度重视。

1987年中国军事科学院成功地研制出“科译 1号”,这标志着我国机译系统从无到有。

1992年,中科院计算机研究所推出了在工作站上运行的“863智能型英汉翻译系统”。

继此之后,随着 PC机的普及和系统操作环境的改善,机器翻译软件逐渐趋向单机操作,并由 DOS版迅速向 Window版过渡。这时的翻译软件主要有天津大通通译计算机软件研究所的“通译”翻译软件、中国软件总公司的“译星”翻译系统和中国科学院语言研究所的“高立”翻译系统等。

20世纪末计算机辅助翻译(CAT: Computer-Aided Translation)随之诞生并高速发展。在此发展阶段,最具代表性的 CAT软件有:国际商用机器公司(IBM)系列计算机辅助翻译软件和国内的雅信 CAT翻译系统等。

面向数据处理(DOP: Data-Oriented Processing)理论系近年来指导机器翻译和计算机辅助翻译的主流理论。该理论最早是由美国计算机专家在 1990年提出,1996年开始受到国际计算机语言界的普遍重视。面向数据处理理论在应用于指导机器翻译和计算机辅助翻译的短短几年间,已外延出以下两项思路或宗旨截然不同的分支理论:

1. 广义语法模式理论(Generalized Grammatical Mode Theory)

该理论旨在基于传统语法分析和实用语法,将原语(source language)和译语(target language)转换过程中对应率较高的句型或表达式结构化和模式化,并编辑成若干语法模式指令来操作机译。这就是所谓的广义语法模式理论。

2. 语言经验模拟理论(Language Experience Simulation Theory)

该理论的基本指导思想是:人类对语言的领悟和创造依赖于以往具体的语言经验,而不是依赖于抽象的语法规则。这一理论旨在利用强大的计算机记忆功能,自动记忆用户的翻译结果或输入的翻译译例(这样记忆的翻译结果或输入的翻译译例既可以是整句形式又可以是句子片段即短语形式)。随

[收稿日期]2001-11-05

[作者简介]李鲁(1957-),男,东南大学外语系教授,硕士生导师,研究方向:科技英语,科技翻译,机器翻译。

之,系统建立数据记忆库,来模拟“语言经验”,或者说是积累“语言经验”。翻译过程中,系统借助高速计算,对需要翻译的内容进行快速分析和对比,并通过高速搜索引擎搜索各类数据记忆库,瞬间找出相同或相似的句子或句子片段,进行匹配机译。

略作分析对比不难看出:语言经验模拟理论实际上是从第一语言习得(First Language Acquisition),也就是说从母语习得的角度,来透视语言的输入和输出。按该理论的基本指导思想来解释,也就是说,就像我们理解和输出母语那样,我们并不是依赖于抽象的语法规则,而是依赖于以往我们在母语语言环境中积累的大量的具体的语言经验。而这种依赖往往是以母语语感的形式,潜意识地自然而然地运作。因此,可以说,语言经验模拟理论旨在模拟第一语言习得的输入和输出过程。反之,基于传统语法分析和抽象语法规则的广义语法模式理论,模拟的则是第二语言学习(Second Language Learning)的非自然的输入和输出过程。也就是说,学习第二语言的人,由于语言环境限制,往往需要依赖语法分析和抽象的语法规则来理解或输出语言。广义语法模式理论模拟的就是这一过程。

上述两种理论似乎出于两种完全不同的思路,但在近年来最新开发的计算机辅助翻译系统中,这两种理论却往往融为一体,相辅相成(这些系统如国际上的 IBM系列计算机辅助翻译系统,国内如雅信 CAT计算机辅助翻译系统等等)。在先进的计算机辅助翻译系统中,语言经验模拟理论往往既体现于该系统的计算机辅助翻译功能上,又体现于系统机译(即系统自动翻译)功能上,而广义语法模式理论则往往体现于并主导后者功能。显而易见,单靠经验数据记忆库,或单凭广义语法模式,都不能使机器翻译或计算机辅助翻译理想化;而两者的结合应用则的确使我们朝着这个方向迈出了一大步。

二、国内英汉翻译系统机译质量现状

综观国内目前流行的一些主要的机器翻译或计算机辅助翻译软件,如大通通译计算机软件研究所的“通译”、中软总公司的“译星”、中科院语言研究所的“高立”和雅信 CAT计辅翻译系统等(除这些之外,当然还有一些更为大众化的软件,如“金山快译”、“东方快车”等),就这些翻译软件目前的机译质量来看,仍未达到令人满意的程度。而且,这些软件的机译错误具有一定的共性。且看下列机译结果(其中的译文选取的是雅信 CAT系统的机译结果;用其它翻译系统机译时,机译结果整体上都十分相似,即大同小异,反映的基本上是同样的问题):

1) It is clear that they have to compete with IBM and some other well- known companies at all costs.

很明显他们不得不与国际商用机器公司和其它的著名的公司不惜任何代价竞争。

2) It is clear that they have to compete with IBM and some other well- known companies at all costs if they decide to enter the computer market.

很明显他们不得不与国际商业机器公司和其它的著名的公司不惜任何代价如果他们决定进入计算机市场竞争。

3) Gases differ from solids because the former have greater compressibility.

气体不同于固体因为前者有更大的压缩性。

4) In addition, the advance of technology cannot be understood without consideration of its interplay with social, economic and psychological forces.

另外,那进步的技术不能了解不考虑的它的相互影响同社会的,经济的和心理学的作用力。

分析对比以上四例不难看出,目前国内新近开发的一些英汉翻译系统在机译英汉两语语序相近或较简单的英文语句时还算准确或至少可以接受(见例 1例 3),具有一定的利用价值,其机译结果不做改动或略做改动即可成型;然而,当机译英汉两语语序有所不同或句型略复杂(如复合句等)的英语句子时,几乎无准确度可言(见例 2例 4),即其利用率极低或无利用率。在后一情况下,译者一般宁愿舍弃自动机译,而借助计算机辅助翻译功能利用键盘或鼠标手动选词组句。实际上,上述四例选取的仅仅是比较简单或结构较规则的句子,旨在显示一些最基本的机译问题。

除此之外,根据笔者所做的应用测试来看,国内目前开发的英汉翻译系统仍无法正确机译若干类型的英文语句,如部分否定句、否定转移句、隐含反义句等。

三、纠正翻译系统机译错误初探

笔者认为,提高计算机辅助翻译速度,除了积累例句和定义常用短语之外,更重要的是要提高系统机译的质量。随着机译质量的提高,机译利用率将必然上升,这样就可以大大减少译者手动选词组句所耗费的时间,使译者能够在大多数情况下基于系统

机译结果,只要略做改动或调整,即可完成全句翻译[①]。

那么如何来提高机器翻译的质量呢?实际上,解决这个问题的一个有效方法,就是从大多数翻译软件所采用的广义语法模式体系入手。大多数翻译软件已随系统提供了约 2万到 5万条广义语法规则来操作机器翻译,但由上节分析所示,其机译质量远不理想。因此,要使系统机译具有利用价值,我们有必要通过添加或修改某些广义语法模式定义指令,来纠正机译错误,以提高机译准确度。

1. 纠正机译语序错误

首先,我们必须解决机器翻译中最容易产生的一种译句语序混乱的现象。通过查看系统机译语法模式并比较上文例 1、例 2不难看出,造成例 2译句语序混乱的原因是 complete with的语法模式释义: compete with* \与# 1竞争。由于在此汉译释义中,原文替换部分"* "的位置有所改变,即"# 1"被移至"竞争"之前,这样一来,虽说在机译简单句时一般不会出现什么问题(见例 1),但一旦原文替换部分"* "后面跟上一个无标点分开的连词从句时,便会导致译句语序混乱(见例 2)。为了解决这一问题,我们可以对 compete with这一表达式进行扩展定义,添加其扩展型广义语法模式;如在雅信 CAT等翻译系统中可打开系统语法库,按指令书写格式添加以下扩展型广义语法模式:

aa* compete with* {con}* ^ ^ @ # 3,# 1与# 2竞争

以对 compete with短语进行连词扩展定义,并进行相应的中文释义定义和语序调整定义。添加后,便可得到此机译结果: "如果他们决定进入计算机市场,很明显他们不得不与国际商业机器公司和其它一些著名的公司不惜任何代价竞争。"这样也就纠正了添加前的机译错误,清除了语序混乱现象。实际上,此类问题均可按这种添加扩展语法定义的方法解决。又如(以下译例中,译文 a为添加前机译结果,译文 b为添加后机译结果):

5) He still had to apologize to her though he thought she was rather rude.

a. 他仍不得不向她虽然他认为她是相当粗鲁道歉。

b. 虽然他认为她是相当粗鲁,他仍不得不向她道歉。(添加模式: aa* apologize to* {con}* ^ ^ @ # 3,# 1向# 2道歉)

6) They are apt to fall out over trivial things since they lived together in London.

a. 他们往往为小事自从他们同居在伦敦吵架。

b. 自从他们同居在伦敦,他们往往为小事吵架。(添加模式: aa* fall out over* {con}* ^ ^ @ # 3,# 1为# 2吵架)

同样,我们也可以通过添加或修改某些语法定义模式,来解决系统机译中存在的其它一些问题。

2. 纠正部分否定句机译错误

纠正部分否定句机译错误,可通过添加"all… not","both… not","every… not"等语法模式系列并进行部分否定释义定义来实现。在这些模式系列中各有一组类似的语法项目[②],以力求尽可能覆盖各种不同的情况(如不同的语态、情态、时态或句式结构等等)。因此,为了方便起见,添加时可直接打开系统语法库文件,通过"复制"来批量添加各模式系列中各组语法项目,然后分别略做改动即可。最后,再进行"批量更新"。

由于篇幅有限,现仅举以下四句为例(译文 a为添加前机译结果,译文 b为添加后机译结果):

7) All the dogs here won't bite passers-by.

a. 全部这里的狗不会咬过路人。(误译为全部否定)

b. 并非所有这里的狗都会咬过路人。(添加模式: aaall* will not* ^ (400)^并非所有# 1都会# 2)

8) All applicants cannot go abroad.

a. 全部报名者不能出国。(误译为全部否定)

b. 并非所有报名者都能出国。(添加模式: aaall* can not* ^ (400)^并非所有# 1都能# 2)

9) It turned out that they both did not keep their words.

① 如只需点击一下鼠标,在系统交互区释义框中稍许调整一下个别词的释义即可。为了便于演示,本文出示的译例机译结果中有几句已对其中个别词的释义做了这样的调整。

② 如"all… not"模式系列的项目组为:

aaall* be not* ^ (400)^并非所有# 1都是# 2
aaall* do not* ^ (400)^并非所有# 1都# 2
aaall* will not* ^ (400)^并非所有# 1都会# 2
aaall* can not* ^ (400)^并非所有# 1都能# 2
aaall be not* ^ (400)^并非一切都是# 2
aaall can not* ^ (400)^并非都能# 2等等。

a. 结果他们两个都不遵守诺言。(误译为全部否定)

b. 结果他们两个并非都遵守诺言。(添加模式: aaboth* do not* ^(400)^两个# 1并非都# 2)

10) When she entered the office, every window was not open.

a. 当她进入那办公室时,所有的窗户没有开着的。(误译为全部否定)

b. 当她进入那办公室时,并非所有的窗户都是开着的。(添加模式: aaevery* be not* ^(400)^并非所有的# 1都是# 2)

3. 纠正否定转移句机译错误

纠正常见否定转移句型的机译错误,可同样按上述批量处理方法,通过添加"not… because of","not… because"等语法模式系列并进行否定转移释义定义来完成。且看以下实例(译文 a为添加前机译结果,译文 b为添加后机译结果):

11) I won't support a candidate because of blackmail and intimidation.

a. 我不会支持一候选人因为勒索和恐吓。

b. 我不会因为勒索和恐吓而支持一候选人。(添加模式: ddwill not* because of* ^(400)^不会因为# 2而# 1)

12) The airplane did not crash because of engine failure.

a. 那飞机没有坠毁因为发动机故障。

b. 那飞机不是因为发动机故障而坠毁。(添加模式: aado not* because of* ^(400)^不是因为# 2而# 1)

13) In this small town, they never suffered discrimination because they were Jews.

a. 在这个小城镇,他们从未遭受岐视因为他们是犹太人。

b. 在这个小城镇,他们从未因为他们是犹太人而遭受岐视。(添加模式: aanever* because* ^(400)^从未因为# 2而# 1)

4. 纠正隐含反义句机译错误

纠正此类机译错误,一般只要给系统添加以下两项语法模式并进行隐含反义释义定义即可:

aait be {art} {adj}* that have no* ^(100)^# 1再@ 2也有# 2

aait be {art} {adj}* that never* ^(100)^# 1再@ 2也会# 2①

且看实例(译文 a为添加前机译结果,译文 b为添加后机译结果):

14) It is a long lane that has no end.

a. 没有尽头是一长小路。

b. 小路再长也有尽头。(即指:凡事必有转机。)

15) It is an ugly wife that has no husband.

a. 没有丈夫是一丑妻子。

b. 妻子再丑也有丈夫。

16) It is a good horse that never stumbles.

a. 从不失蹄是一好马。

b. 马再好也会失蹄。

17) It is a good workman that never blunders.

a. 从不失误是一好工匠。

b. 工匠再好也会失误。(即指:智者千虑必有一失。)

由于上述这两类隐含反义句具有相当特别且极为规则的句型结构(见上文添加的两个语法模式),因此系统机译时一般不会将普通的先行主语句或强调句与其相混淆。且看以下机译结果:

18) It is quite likely that John wronged her.

约翰冤枉她是相当有可能的。

19) It is the force of gravity that makes heavy things fall toward the ground.

正是地心引力使重物落向地面。

以上只是笔者所作的一些初步尝试,触及的只是冰山一角。然而,除上述翻译系统语法库所存在的问题之外,在目前国内的翻译软件中翻译系统对于某些词项的词义定义也值得推敲,有些显然不够准确或合理,尤其是某些短语的定义影响了系统语法库中某些语法模式的运作。另外,影响机译语法模式运作的还有词性和属性的定义。据笔者查看,在上述提到的这些翻译软件的系统词库中仍有相当一部分词汇(其中包括一些常用词)未作词性或属性定义,或在一词多词性的情况下词性定义不全;机译时若遇到这些词,那么含有词性或属性定义的语法模式便无法运作。无疑,解决上述这些问题理应为翻译软件开发者改进和完善系统语法库和系统词库的一个最为重要的部分。同时,我们也期待翻译软件开发者能不断在技术上做出进一步的突破,如突破语法模式定义中的替换部分只能有 3个的限数,以使准确

① 由于雅信 CAT2.5系统词库中遗漏了对"a"一词的冠词词性的定义("an"和"the"的冠词词性均已定义),因此需要补上(即在其词性栏中添上"art")后,才能使这两个语法模式在该系统中遇到冠词"a"时也能正常运作。

调整复杂句型的机译译文语序成为可能 ①

总之,尽管计辅翻译与机器翻译目前仍问题重重,但其前景不可估量。毫无疑问,我们告别繁琐的翻译劳动的时代即将到来!

A probe into machine translation and computer-aided translation

LI Lu

(*Foreign Languages Department*, *Southeast University*, *Nanjing* 210096, *China*)

Key words machine translation (MT); computer-aided translation (CAT); grammatical modes (GM); data-oriented processing (DOP)

Abstract The research of machine translation (MT) and computer-aided translation(CAT) have made some progress in 20th century, But many MT errors are still be made by the existing domestic E-C translation systems. An effective solution to these errors is adding some neessary grammatical modes to the system.

(上接第 82页)

Technology innovation and property rights

ZHANG Zong-qing

(*Department of Economy*, *Southeast University*, *Nanjing* 210096, *China*)

Key words technology innovation; intellectual property rights; spill-over effect; innovation inspiritment; innovation ability

Abstract On the one hand, intellectual property rights protect proplé's innovation capital against being dissipated and thus offer innovation inspiritment; on the other hand, intellectual property rights provide people with innovation opportunities and thus prompt the spread and application of technology. The utilization of opportunities, however, depends on other inspiritment mechanisms and also on peoplé's ability to make use of those opportunities.

① 本文主要参考文献:《中国翻译》,《中国科技翻译》等杂志。