2022年11月 外国语 November 2022
第45卷第6期 Journal of Foreign Languages Vol.45 No.6
文章编号:1004-5139(2022)06-0077-11 中图分类号:H059 文献标识码:A

机器翻译与人工翻译相辅相成

冯志伟[1,2],张灯柯[1]
(1.新疆大学,新疆 乌鲁木齐 830046;2.大连海事大学,辽宁 大连 116026)

摘 要:本文介绍了基于规则的机器翻译、统计机器翻译和神经机器翻译的发展历程,阐释机器翻译研究需要有语言学知识和常识的支持,主张不要过分地迷信目前广为流行的基于语言大数据的连接主义方法,不要轻易地忽视目前受到冷落的基于语言规则与常识的符号主义方法,应当把基于语言大数据的连接主义方法和基于语言规则与常识的符号主义方法巧妙、精准地结合起来,把机器翻译研究推向深入。本文指出,机器翻译将成为人工翻译的好朋友和得力助手,机器翻译和人工翻译应当和谐共生,相得益彰。
关键词:机器翻译;人工翻译;符号主义;连接主义;神经网络;深度学习

Machine Translation and Human Translation Boost Each Other

FENG Zhiwei[1,2], *ZHANG Dengke*[1]
(1. Xinjiang University, Urumqi 830046, China; 2. Dalian Maritime University, Dalian 116026, China)

Abstract: This paper introduces the development of rule-based machine translation, statistical machine translation and neural machine translation, and illustrates that machine translation research needs to be supported by linguistic knowledge and common sense, and proposes that the currently widely popular connectionist approach based on linguistic big data should not be overly fetishized, and the symbolist approach based on linguistic rules and common sense, which is currently receiving a lukewarm reception, should not be easily ignored. The connectionist approach based on linguistic big data and the symbolist approach based on linguistic rules and common sense should be skillfully and precisely combined, so as to push the research of machine translation into depth. This paper points out that machine translation will become a good friend and powerful assistant of human translation, and that machine translation and human translation should boost each other.
Key words: machine translation; human translation; symbolism; connectionism; neural network; deep learning

近年来,由于神经机器翻译的成功,它的翻译能力往往容易被夸大,对于机器翻译技术夸大宣传的声音不绝于耳。本文回顾了基于规则的机器翻译、统计机器翻译和神经机器翻译的发展历程,说明了机器翻译研究需要有语言学知识和常识的支持,目前机器翻译缺乏翻译所需要的"人文硬核",主张机器翻译和人工翻译和谐共生,相得益彰。

1.基于规则的机器翻译

机器翻译(Machine Translation,简称MT)是使用计算机进行跨语言自动翻译的一个学科,也是人工智能研究的一个重要领域(冯志伟2018:35-48)。

1947 年 3 月 4 日,在世界上第一台电子计算机研制出来不久,美国洛克菲勒基金会自然科学部主任韦弗(W. Weaver)给控制论(cybernetics)的奠基人维纳(N. Wiener)写过一封信,提出研制机器翻译的想法,试探维纳这位学界泰斗对于机器翻译的看法。韦弗在信中写道:"我怀疑是否真的建造不出一部能够做翻译的计算机?即使只能翻译科学性的文章(在语义上问题较少),或者翻译出来的结果不怎么优雅(但能够理解),对我而言,都值得一试。"维纳是一位严肃的学者,他对翻译的复杂性有深刻理解,经过一个月的反复思考,他在 4 月 30 日给韦弗的回信中直言不讳地说:"老实说,恐怕每一种语言的词汇,范围都相当模糊;而其中表示的情感和言外之意,要以类似于机器翻译的方法来处理,恐怕不是很乐观的吧!"

维纳的回答给韦弗泼了一瓢冷水。但是,韦弗仍然坚持自己的观点,他于 1949 年在题为《翻译》(*Translation*)的备忘录中,明确提出了关于机器翻译的建议,认为可以使用解读密码的方法进行机器翻译(Weaver 1955: 15 – 23)。

1947 年,英国数学家图灵(Alan Turing)在写给英国国家物理实验室的一份报告中谈到他建造计算机的计划时提出了机器翻译的主张。他指出,翻译是一种智能活动,与人的智能有关,因此,机器翻译可以显示计算机的"智能"。1954 年 1 月 7 日,美国乔治城大学(Georgetown University)在 IBM 公司支持下,进行了世界上第一次机器翻译实验,使用 IBM – 701 计算机把 60 个俄语句子自动翻译成英文(Dostert 1955: 124 – 135)。1954 年 1 月 8 日美国《纽约时报》(*New York Times*)报道了这个实验,引起新闻媒体的极大关注(Hutchins 1997: 192 – 252)。第一次机器翻译实验的成功激发了人们对于机器翻译的热情。尔后十年,机器翻译研究开花结果,并且有了很大的进展。苏联、英国、中国、日本都相继开展了机器翻译的研究工作(冯志伟 2018: 35 – 48)。

著名数理逻辑学家巴希勒(Y. Bar-Hillel)也参加了机器翻译的研究。不过,他清醒地认识到,从当时的技术条件和语言学研究水平来看,全自动高质量的机器翻译(Full Automatic and High Quality Machine Translation, 简称 FAHQMT)是难以实现的。他在《语言自动翻译现状》一文中举了一个很有意思的例子来说明机器翻译的困难性:Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy. (小约翰在寻找他的玩具盒子。他最后找到了它。那个盒子在游戏围栏里。他非常高兴。)巴希勒指出,在这段短文中的单词 pen 有两个意思:一个意思是"钢笔",一个意思是"游戏围栏"。计算机要把这段短文中的单词 pen 正确地理解成"游戏围栏"而不能理解为"钢笔",是一件不容易的事情。因此,巴希勒断言,要对这段短文进行正确的机器翻译是不可能的(Bar-Hillel 1960: 91 – 163)。

学术界和实业界人士一致认为,由于机器翻译的复杂性和困难性,应当对当时的机器翻译水平认真进行评测。随后他们逐步取得共识,这种共识导致美国成立了"语言自动处理咨询委员会"(Automatic Language Processing Advisory Committee,简称 ALPAC),并于 1966 年以美国国家科学院的名义正式发布了 ALPAC 报告,该报告对于机器翻译提出了批评性意见,致使美国政府切断了对于机器翻译研究的资金资助(冯志伟 2018: 35 – 48)。

不过,有些机器翻译的开发者并没有灰心丧气,他们仍继续坚持机器翻译研究。在加拿大,山蒂奥(Chandioux)研制了像 METEO 这样的实用性机器翻译系统,可以进行关于天气预报的英语到法语的机器翻译(Chandioux 1976: 127 – 133)。在实业界,也研制了像美国的 SYSTRAN 这样的商用机器翻译系统。

法国格勒诺布尔理科医科大学的沃古瓦(B. Vauquois)对于机器翻译研究的构架(architecture)进行了概括(见下页图 1)。他把这个时期的机器翻译研究构架总结为三种模式:1)直接翻译模式(Direct Translation model);2)转换模式(Transfer model);3)中间语言模式(Interlingua model)。

1)直接翻译模式:把源语言文本(source text)中的单词直接翻译为目标语言文本(target text)的单词。每一个词条就是一个小程序,担负着翻译相应单词的任务;2)转换模式:对源语言文本进行分析(source language analysis),然后使用规则把表示源语言文本语义/句法结构(semantic/syntactic structure)的分析树(parsing tree)转换为表示目标语言文本的语义/句法结构的分析树,再从目标语言的分析树进行目标语言生成(target language generation);3)中间语言模式:把源语言的句子分析为某种抽象的意义表达式,叫作中间语言,然后从中间语言表达式生成目标语言。沃古瓦把这三种模式使用一个三角形来表示,叫作"沃古瓦三角形"(Vauquois Triangle)。

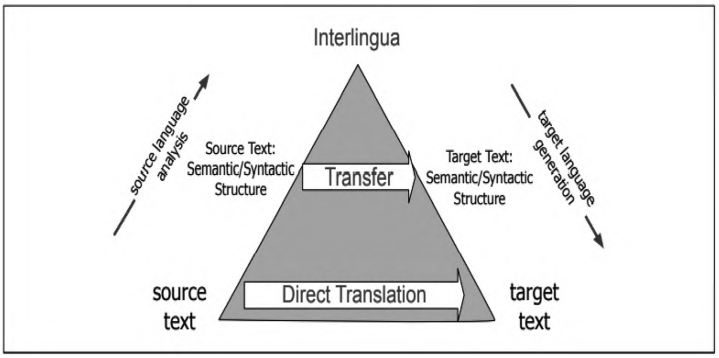

图1　沃古瓦三角形

我们从沃古瓦三角形中可以看出,在机器翻译中,从下层的直接翻译模式,通过转换模式,再到上层的中间语言模式,分析和生成所要求的处理深度是不断增加的,同时,我们还可以看出,从沃古瓦三角形的底部向上移动时,所需要的转换量是逐渐减少的,处理的程度越深,所需要的转换量就越小。因此,在直接翻译模式中,转换量是很大的,几乎每一个单词的知识都是转换需要的知识,要对每一个单词逐一地进行转换。在转换模式中,转换时只需要分析树和语义角色(semantic role)知识。而在中间语言模式中,几乎不需要特定的转换知识(Vauquois 1968: 254 - 260)。

冯志伟于1978 - 1981年间曾经在法国格勒诺布尔理科医科大学应用数学研究所学习机器翻译,沃古瓦就是他的导师。根据沃古瓦三角形和汉语的特点,冯志伟提出了多叉多标记树模型(Multiple-branched and Multiple-labelled Tree model,简称MMT模型),研制了多语言机器翻译系统"法吉拉"(FAJRA),把12篇短文的108个汉语句子自动地翻译成法语、英语、日语、俄语和德语五种外国语,这是世界上第一个把汉语翻译成多种外语的机器翻译系统,研究成果于1982年在布拉格的国际计算语言学会议上用法语发表,于1983年在香港的东南亚电脑国际会议上用英语发表。

以上机器翻译系统都是建立在语言规则和机器词典基础之上的,采用的是符号主义(symbolism)方法,机器翻译的研究者必须对语言规则进行形式化的描述,提取语言特征,建立大规模的机器词典,描述规则和编制词典是一项规模宏大的"语言特征工程"(language feature engineering),费时又费力,这些系统都是基于规则的机器翻译(Rule-Based Machine Translation,简称RBMT)系统(冯志伟、程勇 2020: 55 - 59)。

由于自然语言极为复杂,既涉及语言内部的规则,还涉及语言外部的各种常识,用有限的规则难以穷尽地描述千变万化的语言现象和语言外部常识,基于规则的机器翻译尽管在有限规模的"子语言"(sub-language)中取得局部的成功,但是一旦扩大语言的规模或种类,就往往显得捉襟见肘,难以对付,机器翻译的忠实性(fidelity)和流畅性(fluency)都不理想,翻译的质量不高。

2. 统计机器翻译

20世纪90年代在机器翻译中开始使用统计方法,出现了统计机器翻译(Statistical Machine

Translation,简称 SMT)。由于大规模语料库(large-scale corpus)的发展,建成了像“汉莎尔德”(Hansard)这样的加拿大议会会议录语料库,汉莎尔德语料库中包含英语和法语的双语语料,有助于研究者使用统计方法,从大规模的、真实的语料库中提取语言特征知识(冯志伟 2015:546 – 554)。

由于互联网的发展,机器翻译研究者也可以从互联网网页上获得大规模的、真实的语料,语料库成了统计机器翻译的重要知识来源。统计机器翻译研究者可以运用单词本身信息或句子长度信息等简单的提示信息,从双语语料库中抽取对齐的双语句子,获取丰富的语言特征知识。语料库和统计方法的使用明显提高了机器翻译的质量(Gale *et al.* 1991:177 – 184)。

与此同时,IBM 公司直接根据语音识别中的“噪声信道模型”(noisy channel model)提出两个相关的统计机器翻译范式。一个范式是布朗(P. F. Brown)等人研制的统计机器翻译生成算法,叫作 IBM 模型(包括 1 – 5 个模型),通过“坎戴德系统”(Candide System)来实现,其生成算法的细节(除了解码器之外)已经公之于众,美国政府对于“坎戴德系统”给了部分资助,从而有力地支持了这个研究团队的工作,“坎戴德系统”的机器翻译质量超过了基于规则的机器翻译系统(Brown *et al.* 1990:79 – 85)。另一个范式是 IBM 公司的贝尔格(A. Berger)等提出的最大熵算法(Maximum Entropy algorithm,简称 MaxEnt 算法),这是一种逻辑回归算法,可以把各种不同的语言特征分别结合起来(Berger *et al.* 1996:39 – 71)。

2002 年,奥赫(Och)改进了这种算法,提出了分辨训练方法(discriminative training approach),进一步提高了统计机器翻译质量 (Och 2002:295 – 302)。

世纪之交,大多数机器翻译研究都转向采用统计方法。在统计机器翻译中,马尔库(D. Marcu)等学者提出了基于短语的统计机器翻译(Phrase-based SMT)方法,这种方法使用短语偶对(phrase pairs)来提升统计机器翻译的效果 (Marcu *et al.* 2002:133 – 139)。在机器翻译评测中,帕皮内尼(Papineni)等研制了用于评测的指标 BLEU(Bi-Lingual Evaluation Understudy),并使用对数线性模型来优化像 BLEU 这样的评测指标。奥赫还提出了“最小错误率训练”(Minimum Error Rate Training,MERT)的机器翻译评测方法,这类方法来自语音识别模型(Och 2003:160 – 167)。奥赫和奈依(Ney)还开发了 GIZA 这样的统计机器翻译工具包(Och & Ney 2003:19 – 51),科恩(P. Koehn)等又开发了 Moses 这样的统计机器翻译工具包。

在世纪之交还出现了一些基于短语和句法结构的统计机器翻译新方法,这类方法主张把基于统计的经验主义方法与基于规则的符号主义方法结合起来(冯志伟 2015:546 – 554)。例如,基于“转录语法”(transduction grammar)的模式把一个并行的句法树结构指派给语言中的一个句子偶对,对句法树进行调整词序的操作,就可以改善目标语言句子的译文。从生成的角度来看,我们可以把转录语法看成是在两种语言中生成对齐的句子偶对(sentence pair)的方法,在这方面使用得最广泛的转录语法有“反向转录语法”(inversion transduction grammar)和“同步上下文无关语法”(synchronous context-free grammar) (Chiang 2005:263 – 270)。这些统计机器翻译系统尽管使用了统计方法,但是仍然关注着基于规则的符号主义方法。

3. 神经机器翻译

自 2006 年以来,统计机器翻译进一步发展成神经机器翻译(Neural Machine Translation,简称 NMT),神经机器翻译也是基于双语或多语并行语料库数据的(冯志伟 2021:87 – 100),这类机器翻译完全是经验主义的,神经机器翻译系统根据双语语料库进行深度学习,就可实现机器翻译,不再需要规模宏大而艰巨的“语言特征工程”,几乎完全抛弃了基于语言规则的符号主义方法。

例如，在德英机器翻译中，如果输入德语句子 Er wollte nie irgendeiner Art von Auseinanderersetzung teilnehmen（他不打算参与任何种类的争论），神经机器翻译系统可以生成 n 个最好的候选英语句子译文，称为“n-best 句子列表”（图 2）：

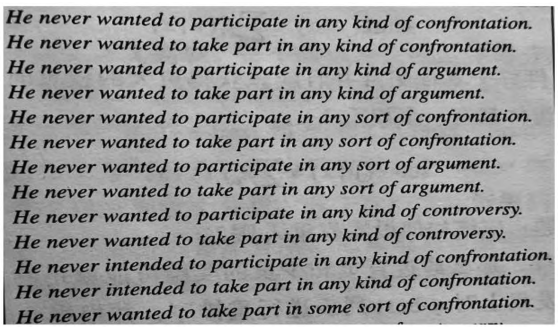

图 2　n-best 句子列表

系统从 n-best 句子列表中，根据对数概率等因素，选出概率最大的句子 He never wanted to participate in any kind of confrontation 作为译文输出。

神经机器翻译把自然语言中的单词符号映射到 N 维空间中，表示为“词向量”（word vector），如图 3 所示（冯志伟 2019：1－10）。

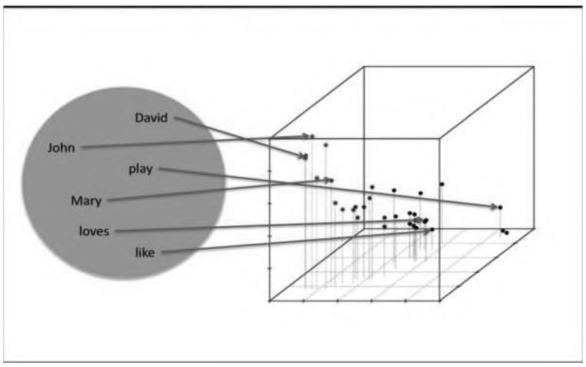

图 3　N 维空间中的词向量

在图 3 中，David、John、Mary、play、loves、like 等离散的单词符号，都被映射到“向量空间”（vector space）中，成了不同的词向量。词向量是单词语义的一种新的表达方式，单词的语义不是通过它指称的事物来表示的，而是通过它所处的上下文的“分布”来表示的，这样的语义研究，叫作“向量语义学”（vector semantics）。

在向量空间中，意义相近的单词位置彼此靠近，如下页图 4 所示向量空间中的 developing、growing 等单词聚集在相近位置，这是因为它们在实际语言中所处的上下文相似；having、had、have 等单词聚集在另一个相近的位置，这是因为它们的语法功能相近；而意义与它们不同的 right、left 等单词则聚集在另一个位置，这是因为它们在实际语言中所处的上下文与 developing、growing 以及 having、had、have 等单词不同。下页图 4 是计算机自动生成的向量空间。

根据向量语义学，神经机器翻译不需要对于离散的语言符号进行计算，而只要把离散的语言符号转换为词向量嵌入到向量空间中进行计算，整个计算是针对没有语言符号的实数值（real values）进行的；而基于语言规则的机器翻译和统计机器翻译则需要对于语言符号及其特征表示

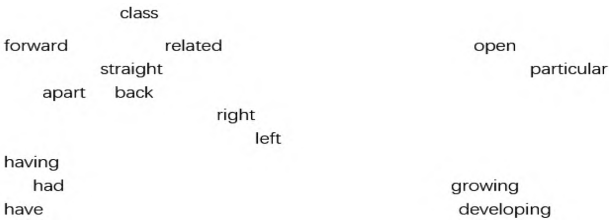

图 4　意义相近单词在向量空间中的分布

(features representations)进行描述和计算,这是非常艰巨的语言特征工程。在神经机器翻译中,由于把离散的单词符号都映射为向量空间中的词向量,不需要规模巨大的语言特征工程,也不需要手工设计语言特征,计算机能够自动从双语语料库中获取到语言特征,并对语言特征进行计算。神经机器翻译抛弃了传统的基于规则的符号主义方法,从符号主义转向了连接主义。这是21 世纪机器翻译研究中一次具有战略意义的重大转移 (冯志伟 2021: 87 – 100)。

其实,使用基于大数据连接主义方法的神经网络来进行机器翻译的思想在世纪之交就有人提出过了。神经网络曾经在不同时间应用于机器翻译的不同方面。例如,早在 2006 年,史维克等(Schwenk *et al.*)就提出,在基于 IBM 模型的西班牙语 – 英语的统计机器翻译中,可以使用神经网络语言模型来代替 n – 元语法模型(Schwenk *et al.* 2006: 723 – 730)。这是神经机器翻译的萌芽。

从 2013 年开始,神经机器翻译就逐渐取代了统计机器翻译。2013 年,卡尔士布莱尔(N. Kalchbnner)和布伦松(P. Blunsom)提出“循环连续的翻译模型”(recurrent continuous translation models),使用“循环神经网络解码器”(Recurrent Neural Network decoder)方法来进行机器翻译。

2014 年,巴丹诺(D. Bahdanau)、本吉奥(Y. Bengio)等提出使用“编码器 – 解码器模型”(encoder-decoder model)来学习短语表示,把“编码器 – 解码器”这样的神经网络模型引入到自然语言处理中来,苏慈凯维(I. Sutskever)等说明了怎样使用神经网络来进行“序列到序列”(sequence to sequence)的机器学习。

对于输入进行“软加权”(soft weighting)生成解码器的思想(这是注意力机制的中心思想)是格拉维斯(A. Graves) 在研制手写体字符识别中首先提出的。2015 年,巴丹诺发展了这样的思想,把它命名为“注意力机制” (Attention mechanism),并应用于机器翻译中。2017 年, 法斯瓦尼(A. Vaswani)等提出了基于注意力机制的神经机器翻译模型。从此,神经机器翻译成了机器翻译的主流 (Koehn 2020: 24 – 31)。

编码器 – 解码器网络(encoder-decoder network)又叫作序列到序列网络(sequence-sequence network),这是一种神经网络语言模型,这个模型能够根据上下文生成任意长度的、恰当的输出序列,非常适合用来做机器翻译。下面图 5 是编码器 – 解码器构架的图示。

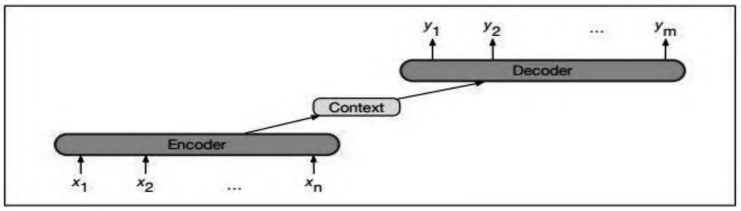

图 5　编码器 – 解码器构架

编码器 – 解码器构架包括三个组成部分:1)一个编码器(Encoder),它接受输入序列 x_1^n(x_1, x_2, …, x_n),生成相应的上下文序列。2)一个上下文向量,记为 Context,它是上下文序列的函数,上

下文向量把输入的精粹内容以词向量形式传给解码器。3)一个解码器(Decoder),它接受上下文向量作为输入,生成隐藏状态的任意长度序列,由此得到输出状态的相应序列 $y_1^m(y_1, y_2, \cdots, y_m)$。

把图5中的编码器-解码器构架与图1中的沃古瓦三角形相比较可以看出,这类编码器-解码器构架很像沃古瓦三角形中的中间语言模型,上下文向量就相当于沃古瓦三角形中的中间语言。这样的机制使单词和单词上下文向量表示能够直接被解码器访问,生成目标语言。

语言数据是神经机器翻译最重要的资源,神经机器翻译中的语言知识都是从语言数据中获得的。2020年,OpenAI公司的卡普兰(Jared Kaplan)和他的合作者提出,神经网络语言模型遵循着"数据升级定律"(data scaling laws)。这个定律说明,向神经网络输入的数据越多,这些网络的表现就越好。这意味着,如果收集更多的数据,并在越来越大的范围内进行深度学习,神经网络语言模型可以做得越来越好。因此,我们应当加强神经机器翻译中的语言数据资源的获取和研究。最近在自然语言处理研究中提出的"预训练+微调"范式,用大规模的数据进行预训练,用小量的数据在下游进行微调,将有助于缓解语言数据资源匮乏的困难(冯志伟、李颖 2021:1-14)。

目前,神经机器翻译已进入大规模实用阶段,在英汉和汉英机器翻译方面,通用文本的翻译正确率已经超过90%,可满足日常生活、新闻报道、海外旅游、商品说明书、旅馆预订服务、交通信息咨询、天气预报查询等一般文本翻译的需要。如果翻译任务的风险不高,翻译结果可选,那么神经机器翻译是有非常广阔的市场的。由于神经机器翻译速度快、容量大,受到了用户的欢迎。如果使用译后编辑(Post Editing,简称PE),对神经机器翻译的译文进一步修改或润色,效果还会更好。神经机器翻译正在一步步地向人工翻译水平逼近。机器翻译已经从人类的梦想成了现实。

4. 机器翻译技术缺乏人文硬核

我们认为,目前广为流行的神经机器翻译尽管已经取得了很大进步,具有一定模拟人类语言内部结构的能力,但是模拟外在世界以及社会历史背景的能力还十分有限。从本质上说,神经机器翻译具备的智能还不是完善的人类智能,只是初级阶段的人类智能。所以,冯志伟(2018:35-48)曾经表示,目前的自然语言处理还处于初级阶段。

翻译是人类的高级智能活动,翻译活动不仅涉及语言内部的结构,还涉及语言外部的日常生活知识、社会知识、历史知识、文化背景知识、人们的心理状态、人们的情感愿望等极为复杂、丰富多彩的因素,这些因素构成了翻译的"人文硬核"(humanity core)。以模式识别技术为基础的神经机器翻译难以处理这些复杂而丰富的"人文硬核",因此,当神经机器翻译遇到"人文硬核"时,就往往会捉襟见肘、左右为难。

就在几个月前,我们使用目前最受欢迎的神经机器翻译系统DeepL来翻译巴希勒1960年提出的那个英语例子:*Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.* 。

当年巴希勒认为,机器翻译不可能把这个例子中的多义词pen正确地翻译成"游戏围栏"。62年之后,神经机器翻译系统DeepL仍然没有这样的翻译能力,这是DeepL给出的译文:小约翰一直在寻找他的玩具箱。最后他找到了。盒子就在笔里。约翰非常高兴。神经机器翻译系统DeepL没有能力解决巴希勒62年前提出的问题,仍然把 *The box was in the pen* 错误地翻译为"盒子就在笔里"。产生这种错误的原因在于:

第一,神经机器翻译系统没有外部的知识,不知道玩游戏需要一个"游戏围栏",这个"游戏围栏"在英语中就叫作pen,因此,就把pen错误地翻译成"笔"。

第二,神经机器翻译系统也不知道"盒子"的体积大于"笔"的体积这样语言之外的常识,因而

翻译成“盒子就在笔里”,违背了普通的常识。

从1960年至今,62年悄然过去了,但是巴希勒提出的问题仍然没有得到解决。凡是遇到这类涉及语言之外的常识问题,神经机器翻译往往都会出错。这是神经机器翻译的软肋。

尽管神经机器翻译取得了很大成绩,但是,由于深度学习和神经网络技术都是建立在模式识别基础之上的,是一种基于大数据的连接主义方法,缺乏可解释性,难以理解翻译中深刻的“人文硬核”,也不具备处理语言外丰富多彩的各种知识的能力。说到底,神经机器翻译还不具有真正的智能。现在神经机器翻译遇到了严重的“人文硬核”瓶颈。传统的“符号处理”(symbol manipulation)技术可以为神经机器翻译提供有关的语言学规则和语言外常识,从而缓解神经机器翻译“人文硬核”的瓶颈问题,在这样的情况下,基于语言大数据的连接主义方法很有必要与基于语言规则和语言外常识的符号主义方法结合起来。

我们衷心地希望下一代学者不要过分地迷信目前广为流行的基于语言大数据的连接主义方法,不要轻易地忽视目前受到冷落的基于语言规则和语言外常识的符号主义方法。我们应当让下一代学者做好创新的准备,积极探索新的研究范式,把基于语言大数据的连接主义方法与基于语言规则和语言外常识的符号主义方法巧妙、精准地结合起来,从而把机器翻译的研究推向深入(冯志伟2021:87-100)。

5. 人工翻译是高级的智能活动

文学作品的翻译充满艰辛却又令人神往,文学翻译与任何其他需要人类丰富创造力的智能活动领域一样,都需要译者具备精湛的人文科学的素养。我们来研究下面摘自18世纪曹雪芹小说《红楼梦》中第四十五回末尾处的一段文字(曹雪芹、高鹗1982:630):

“黛玉自在枕上感念宝钗……

又听见窗外竹梢蕉叶之上,雨声淅沥,清寒透幕,不觉又滴下泪来。”

我们把中文原文用英文词逐个对应如下:

黛玉 自 在 枕 上 感念 宝钗 ……

Dai-yu alone on bed top think-of-with-gratitute Bao-chai ...

又 听见 窗 外 竹 梢 蕉 叶

again listen to window outside bamboo tip plantain leaf of

之上 雨 声 淅沥, 清 寒 透 幕,

On-top rain sound sign drip, clear cold penetrate curtain,

不 觉 又 滴 下 泪 来。

not feeling again fall down tears come .

著名文学翻译家霍克斯(David Hawkes)对于这段文字的英文翻译如下:*As she lay there alone, Dai-yu's thoughts turned to Bao-chai... Then she listened to the insistent rustle of the rain on the bamboos and plantains outside her window. The coldness penetrated the curtains of her bed. Almost without noticing it she had begun to cry.* 霍克斯的译文非常精彩。中文的“枕上”英文读者很难理解,因此,霍克斯只是简单地翻译为she lay there alone。由于中文几乎没有动词时态和语态的变化,因此霍克斯不得不决定将中文“透”翻译为penetrated,而不是was penetrating或had penetrated。霍克斯添加了物主代词her使得her window相比the window更适合那种安静闲适的卧室气氛。为了使不熟悉中国床帷的英文读者能清楚地理解,霍克斯将“幕”翻译为curtains of her bed。最后,短语“竹梢蕉叶”的中文非常优雅,这种四字格短语是有文化品位的标志,但是如果以词对词的方式翻译为英文,那就很糟

糕了,因此霍克斯只是简单地将它翻译为 bamboos and plantains。

显然,这类文学翻译要求译者对源语言中的输入文本的文化背景具备博大精深的理解,同时也需要译者能够老练地、富有诗意地、创造性地运用目标语。

我们使用神经机器翻译 DeepL 对这段文字的翻译结果是:*Daiyu since on the pillow sentimental Baochai ... And heard outside the window above the bamboo scorched leaves, the sound of rain, clear cold through the curtain, do not feel and drops of tears.* 这段机器译文的意思虽然可以大致猜出来,但是文采尽失。

由此可见,将文学作品从一种语言到另外一种语言的高质量翻译,是机器翻译难以胜任的,应当由人工翻译来承担。

在科技翻译中,多义术语的翻译是一个棘手问题。有的多义术语可以表示许多概念。例如,英语的 carrier 至少就代表着如下 12 个概念:1)载波;2)承载子,承载形;3)载体;4)载流子;5)载波频率,基频;6)运载工具,搬运车;7)航空母舰;8)载架;9)带基因者;10)带病体;11)鸡心夹头;12)带菌者。人工翻译时,译者需要根据他丰富而广博的科学知识,从多个不同的术语中选择恰当的术语。而机器翻译很难正确地区分对于像 carrier 这样的多义术语,因此往往会造成翻译的错误。

口语翻译可以分为同声传译和交替传译。同声传译是在不打断说话人讲话的情况下,不间断地把讲话的内容口译给听众的一种翻译方式。交替传译是说话者说完一段话之后,翻译者再口译给听众的一种翻译方式。目前尽管机器翻译也可以做同声传译和交替传译,但是,由于计算机并不理解口译的意义,常常出错。因此,重要的同声传译和交替传译必须由人工翻译来承担。

由此可见,人工翻译是人类高级的智能活动,是机器翻译难以完全替代的。

6. 机器翻译是人工翻译的得力助手

机器翻译是人工智能皇冠上的明珠,它不仅仅需要计算机科学和数学的支持,还需要语言学、心理学、神经科学、社会学、人类学等学科的支持,只有汇聚各种知识,形成强大力量,机器翻译研究才有可能突破“人文硬核”的瓶颈,继续开辟出一片新天地。

复杂的高端翻译工作是计算机无法完全替代的。优秀文学作品的翻译、经典著作的翻译、世界名著的翻译、政策文件的翻译、科学技术文献的翻译、风险性高的翻译、重要的同声传译和交替传译,都是要由人来进行的,由于机器翻译现在还难以理解这些文本中蕴含的丰富多彩的“人文硬核”,一旦翻译失误,后果极为严重。因此,这些高端的翻译工作必须由人来承担。

2016 年在加拿大多伦多举行的一次人工智能会议上,深度学习权威专家辛顿(Geoffrey Hinton)曾说过,“如果你是一名放射科医生,那你的处境就像一只已经在悬崖边缘但还没有往下看的郊狼。”辛顿认为,深度学习非常适合读取核磁共振(MRIs)和 CT 扫描图像,因此,放射科医生已经没有什么用处了,人们应该停止培训放射科医生。

但是,在 6 年后的今天,我们并没有看到哪位放射科医生被取代,成千上万的放射科医生仍然没有失业,他们仍然在辛苦地工作着。人们发现,深度学习在放射学中应用非常困难,至少到目前为止,放射科医生和深度学习二者的优势还处于彼此互补,而不是彼此替代。尽管辛顿因为在深度学习方面的出色研究获得图灵奖,但是实践证明他对于取代放射科医生的预见是错误的,他高估了深度学习的能力。

近年来,由于机器翻译的成功,它的翻译能力往往被夸大了,从传统媒体到新媒体,对于机器翻译技术夸大宣传的声音不绝于耳。几年前我国也有人认为,既然计算机已经可以做机器翻译,

也没必要再培养翻译人员了,外语专业也应当停止招生,翻译人员就要失业了。

据《2019 年中国语言服务行业发展报告》统计:2019 年,全球语言服务产值首次接近 500 亿美元;中国涉及语言服务的在营企业有 360,000 余家。2021 年,中国涉及语言服务的在营企业有 423,547 家,以语言服务为主营业务的在营企业近万家,总产值超过 300 亿元,年增长 3% 以上。在这 300 亿元的产值中,当然也有机器翻译的贡献(据统计,在上述以语言服务为主营业务的上万家在营企业中具有机器翻译和人工智能业务的企业有 252 家)。但是,这其中的绝大部分产值是人工翻译贡献的。因此,人工翻译仍然是我国语言服务产业的主力军。

外语专业仍然在招生。现在,全国开设外语类专业的高校数量多达上千所,其中设立有翻译硕士和翻译本科专业的院校分别有 250 余所和 280 余所,翻译硕士累计招生数达 6 万余人 。

可见,尽管机器翻译在我国有了长足进步,但是,机器翻译并不能代替人工翻译。人工翻译仍有着蓬勃的生命力,还会继续发展。高端的人工翻译是机器翻译永远也取代不了的。机器翻译可以帮助人们获取信息,进行信息沟通和交流,还可以作为自然语言处理系统管道流的一部分,加入自然语言处理的管道式系统中,也可以参与多模态的机器翻译。

在人工智能时代,翻译工作中越来越多地使用计算机翻译技术,如机器翻译译后编辑(machine translation post editing,简称 MTPE)技术、术语数据库(terminology data bank,简称 TDB)技术等,MTPE 可以修正机器译文的错误并润色机器译文,TDB 可以修正并统一机器译文的术语,这些技术有助于提高人工翻译的效率。从事人工翻译的翻译工作者应当与时俱进,欢迎这些技术,学习这些技术,掌握这些技术。

在机器翻译中,使用 MTPE 能有效地提升了翻译质量。图 6 比较了 MTPE 的翻译速度与专业译员人工直接翻译的速度,可以看出,MTPE 把人工直接翻译速度提高了 42% 至 131% 。

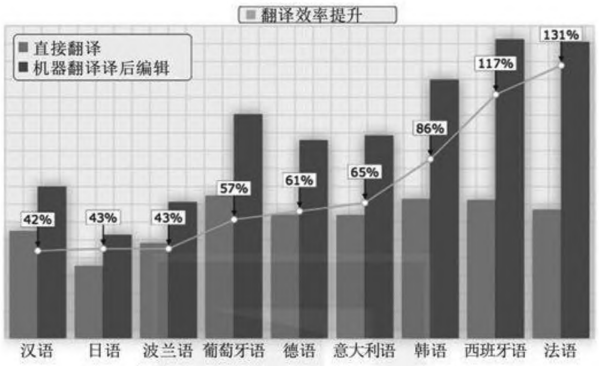

图 6 MTPE 提升翻译效率

我们认为,机器翻译将成为人工翻译的好朋友和得力助手,机器翻译和人工翻译应当和谐共生,相得益彰。

参考文献:

[1] Bar-Hillel, Y. The present status of automatic translation of languages [C] // Alt, F. *Advances in Computers* 1. Academic Press, 1960. 91 – 163.

[2] Berger, A. *et al.* A Maximum Entropy approach to natural language processing [J]. *Computational Linguistics*, 1996, 22(1):39 – 71.

[3] Brown, P. F. *et al.* A statistical approach to machine translation [J]. *Computational Linguistics*, 1990, 16(2): 79 – 85.

[4] Chandioux, J. METEO: Un systeme operationnel pour la tradiction automatique des bulletins meteorologuique destines au grand public [J]. *Meta*, 1976, 21:127 – 133.
[5] Chiang, D. A hierarchical phrase-based model for statistical machine translation [C] //Ann Arbor, MI, *Proceedings for Association of Computational Linguistics* – 2005(ACL – 05):263 – 270, 2005.
[6] Dostert, L. Georgetown-IBM experiment [C] // Locke, W. N. & A. D. Booth. *Machine Translation of Languages*, MIT Press, 1955. 124 – 135.
[7] Gale, W. A. & K. W. Church. A program for aligning sentences in bilingual corpora[C] // *Proceedings for Association of Computational Linguistics* – 1991(ACL – 91), Berkeley, CA, 1991. 177 – 184.
[8] Hutchins, W. J. From first conception to first demonstration: the nascent years of machine translation, 1947 – 1954. A chronology [J]. *Machine Translation*, 1997, 12: 192 – 252.
[9] Koehn, P. *Neural Machine Translation*[M]. Cambridge University Press, 2020. 24 – 31.
[10] Marcu, D. *et al.* A phrase-based, joint probability model for statistical machine translation [C] // *Proceedings for Conference of Empirical Methods for Natural Language Processing* (*EMNLP* 2002). 133 – 139.
[11] Och, F. J. Discriminative training and maximum entropy models for statistical machine translation [C] // *Proceedings for Association of Computational Linguistics* – 2002(ACL – 02), 2002. 295 – 302.
[12] Och, F. J. Maximum error rate training in statistical machine translation[C] // *Proceedings for Association of Computational Linguistics* – 2003(ACL – 03), 2003. 160 – 167.
[13] Och, F. J. & H. Ney. A systematic comparison of various statistical alignment models [J]. *Computational Linguistics*, 2003, 29(1): 19 – 51.
[14] Schwenk, H., Dechelotte, D. & J. -L. Gauvain. Continuous space language models for statistical machine translation[DB/OL] // *Proceedings for Conference of COLING and Association of Computational Linguistics* – 2006 (*COLING/ACL* 2006). Main Conference Poster Sessions. Association for Computational Linguistics, Sydney, Australia, http://www. aclweb. org/anthology/P/P06/P06 – 2093, 2006. 723 – 730.
[15] Vauquois, B. A survey of formal grammars and algorithms for recognition and transformation in machine translation[C] // *IFIP* Congress 1968, Edinburgh, 1968. 254 – 260.
[16] Weaver, W. Translation [C] // Locke, W. N. & A. D. Booth. *Machine Translation of Languages*. Massachusetts: MIT Press, 1955. 15 – 23.
[17] 曹雪芹, 高鹗. 红楼梦[M]. 北京:人民文学出版社,1982.
[18] 冯志伟. 基于短语和句法的统计机器翻译 [J]. 燕山大学学报, 2015, 39(6):546 – 554.
[19] 冯志伟. 机器翻译与人工智能的平行发展[J]. 外国语, 2018, 41(6): 35 – 48.
[20] 冯志伟. 词向量及其在自然语言处理中的应用[J]. 外语电化教学, 2019,(1):1 – 10.
[21] 冯志伟,程勇. 面向汉语自动分析的语言特征工程研究 [J]. 鲁东大学学报,2020, (5):55 – 59.
[22] 冯志伟,李颖. 自然语言处理中的预训练范式[J]. 外语研究,2021, (1): 1 – 14.
[23] 冯志伟. 机器翻译与人工智能[C] // 赵世举,姬东鸿,李佳. 语言学与人工智能跨学科对话. 北京:中国社会科学出版社,2021. 87 – 100.
[24] 冯志伟. 生成词向量的三种方法[J]. 外语电化教学,2021, (1):18 – 26.

收稿日期: 2022 – 08 – 19

作者简介: 冯志伟,研究员,博士生导师。研究方向:计算语言学。
张灯柯,讲师。研究方向:计算语言学。