

目 录

第一章	机器翻译的发展历程与特点	1
1.1	机器翻译的发展历程.....	1
1.2	机器翻译的主要原理与特点.....	2
1.2.1	基于规则的机器翻译.....	2
1.2.2	基于实例的机器翻译.....	3
1.2.3	统计机器翻译.....	4
1.2.4	神经网络机器翻译.....	5
1.3	机器翻译整体特征.....	8
第二章	机器翻译质量评价体系	9
2.1	整体框架.....	10
2.1.1	词汇意义的转换.....	11
2.1.2	短语和句子意义转换.....	12
2.1.3	语篇和语言功能意义转换.....	14
2.2	词汇级评价.....	15
2.2.1	命名实体词.....	15
2.2.2	事实成词.....	19
2.2.3	成语、习语、多字词语.....	20
2.2.4	多义词.....	26
2.2.5	语言处理中的歧义词.....	31
2.3	句子级评价.....	32
2.3.1	短语结构.....	32
2.3.2	句法结构.....	34
2.3.3	句子类型.....	37

2.4	语篇级评价.....	41
2.4.1	内在要素.....	42
2.4.2	语篇功能.....	45
2.5	应用示例.....	49
第三章	机器翻译系统的评测结果与分析.....	51
3.1	词汇级.....	52
3.1.1	命名实体的翻译评测.....	52
3.1.2	数字.....	57
3.1.3	多义词.....	58
3.1.4	成语、习语和多字词语.....	71
3.1.5	专业词汇.....	82
3.1.6	虚词.....	82
3.1.7	联绵词.....	84
3.1.8	组合歧义词.....	85
3.1.9	词汇级评测小结.....	89
3.2	句子级.....	89
3.2.1	短语结构与语序.....	89
3.2.2	句法结构.....	91
3.2.3	不同句型.....	92
3.2.4	句子级评测小结.....	99
3.3	语篇级.....	100
3.3.1	语篇衔接.....	100
3.3.2	语篇功能.....	125
3.3.3	语篇级评测小结.....	129
3.4	机器翻译质量的历时考察.....	129
3.4.1	整体质量打分.....	130
3.4.2	具体分析.....	131
3.4.3	历时评测小结.....	159
第四章	人机关系与翻译教学实践.....	160
4.1	同传的译前准备.....	161

4.1.1	观点调查	162
4.1.2	实验与结果	163
4.1.3	实验后访谈	175
4.1.4	研究结论	179
4.2	译后编辑	180
4.2.1	译后编辑及分类	180
4.2.2	译后编辑的优势和劣势	181
4.3	译前编辑	182
4.3.1	译前编辑的作用	182
4.3.2	译前编辑的效度实验	183
4.3.3	研究结论	188
第五章	总结与展望	189
5.1	机器翻译未来的发展	189
5.2	机器翻译的应用场景	191
参考文献	193
附录一	实体词和数字评测结果示例	196
附录二	英文商用习语评测数据	199

第一章 机器翻译的发展历程与特点

1.1 机器翻译的发展历程

机器翻译（Machine Translation, MT）指利用计算机实现不同语言之间的文本自动翻译的方法或软件系统。机器翻译研究是人工智能研究子领域——自然语言理解（Natural Language Understanding, NLU）下的一个重要方向，也是备受关注的热点和难点课题。

机器翻译的研究工作最早可追溯到 20 世纪四五十年代，在计算机诞生不久就开始了，时至今日已经历了 70 多年的发展历程，期间起起落落，承载了几代人不懈努力和追求的研究历史。机器翻译最初是冷战的产物，很快就吸引了大量的研究人员。但是在 1966 年，美国科学院语言自动处理咨询委员会（Automatic Language Processing Advisory Committee, ALPAC）组织了对英俄机器翻译系统的评测，在那份几十页的评测报告中认定机器翻译研究经费太高，翻译质量太差，根本没有发展的必要，机器翻译研究随后被打入冷宫。机器翻译研究的复苏应该归功于 20 世纪 90 年代统计机器学习的研究成果。基于统计的机器翻译不再依赖语言专家制定的翻译规则。统计学习模型经过大规模双语翻译语料训练，获取翻译的统计“知识”。这种翻译知识可做到与具体的语言无关，从而提高了翻译模型的通用性，机器翻译廉价、快捷的优势逐渐凸显出来。面对广阔的翻译市场，研究者的激情再次被点燃。在随后的 20 年间，统计机器翻译的研究逐步深入，译文质量也不断提升，并逐步迈向了市场。然而，一些机器翻译的核心问题仍然存在，并且随着模型核函数维度的提升，译文稳定性变差，模型泛化能力下降，因此翻译质量提升的空间越来越小。机器翻译研究再一次柳暗花明是 2014 年，一篇使用神经网络

实现机器翻译的论文开启了机器翻译又一个新的时代。到 2019 年，主流机器翻译已经从统计翻译全面转型到神经网络翻译，并不断获得突破性成果。目前性能最好的机器翻译系统都是基于神经网络结构实现的。人们明显感受到了神经网络机器翻译的质量变化，尤其在译文的流利度方面显著超越了基于短语的统计机器翻译，甚至机器翻译质量超越人工翻译的报道也不绝于耳。机器翻译实用化的进程已经全面开启，各种产品（如翻译机）已经走进了人们的日常生活。

1.2 机器翻译的主要原理与特点

从实现原理上划分，机器翻译可分为基于规则的翻译、基于实例的翻译、统计机器翻译和神经网络机器翻译四大类。下面简要阐述各种类型机器翻译的原理与特点。

1.2.1 基于规则的机器翻译

最早机器翻译的构想就是把人类翻译的方式“告诉”计算机，也就是根据语言专家提供的翻译规则和大量的语言知识，转换为机器能够接受的编码形式，再由计算机实现自动翻译。这种翻译方法首先要把翻译规则以一定的形式在计算机内存储，通过检索匹配的翻译规则和双语词典等资源实现翻译。基于规则的机器翻译的原理如图 1-1 所示。

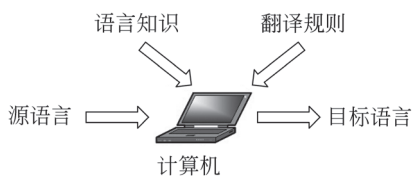


图 1-1 基于规则的机器翻译原理

所谓的翻译规则就是包含了变量的翻译模板。比如：

Rule 1: [X want to DO] → [X 想要 做什么]

模板中有固定的语言成分，也有可变的成分，上例中用 X 表示句子的主语，DO 表示句子的谓语部分。比如，如果英文句子“I want to

borrow the book.”需要译为中文，机器经过分析，认为原文能够利用规则 1 进行翻译，再借助双语词典，就可以得到译文“我想要借书”。

由于语言现象十分复杂，翻译规则繁多且异常复杂，甚至规则之间经常发生冲突，而不符合规则的语言现象也比比皆是，这些都导致基于规则的机器翻译质量不高。再加上人工制定规则的代价很高，基于规则的机器翻译往往只有那些具有较长时间积累的大公司才有实力继续投入经费开展下去。目前商用的、基于规则的机器翻译系统中，比较著名的是 Systran 系统。

基于规则的机器翻译的特点是大量借鉴了人工翻译的方法，其关键技术是语言知识、翻译知识甚至常识如何在计算机内被表示、存储和利用。基于规则的机器翻译对于能匹配规则的情形一般能得到语义准确且流畅的译文，对于不能匹配规则的原文甚至无法得到译文。因此，现在基于规则的机器翻译主要应用在某些受限领域，如天气预报和产品说明书的翻译，因为特定领域的语言变化规则少，相对容易构建基于规则的翻译系统。

1.2.2 基于实例的机器翻译

基于实例的机器翻译源于一种简单的信息组合的思想，即根据已有的人工翻译句对来实现自动翻译。比如，已知如下英汉翻译的句子对：

She bought a new skirt. || 她买了一条新裙子。

He is reading a book. || 他在读书。

机器通过对已有翻译句对的分析，找出语法、词汇和短语等层面的对应关系，自动提取并生成翻译模板。假设现在输入一个新的要翻译的英文句子是“*She is reading a new book.*”，那么机器就可以根据翻译模板，再经过信息重组得到译文“她在读一本新书”。

基于实例的机器翻译相当于通过记忆获取翻译知识，可以节省部分人工成本，其关键技术包括翻译实例的获取、翻译对应单位的划分、对应翻译模板的信息提取和翻译结果的整合等。该翻译方法自 20 世纪 80 年代提出以来，发展至今也面临诸多困难。基于实例的机器翻译给予人们的研究启示是：机器可以从大量翻译实例中基于经验学习翻译方法。

1.2.3 统计机器翻译

统计机器翻译相比前两种翻译方法更偏重于对语言数据的利用和计算，是基于统计数学模型实现的自动翻译。翻译的实现思路借鉴了通信原理中的噪声信道模型：将源语言的句子视作某个目标语的句子通过一个噪声信道受到干扰而产生的。翻译的任务就是根据原文找到没有被干扰的目标语句子 $p(e|f)$ 的过程，即：给定某个原文 f ，找到与其对应的最有可能的译文 e 。用公式描述就是求使条件概率取得最大值的那个译文，如公式 1-1：

$$\text{公式 1-1: } \bar{e} = \arg \max p(e|f) = \arg \max p(e|f) p(e)$$

其中， $p(e|f)$ 表示在给定译文 e 的情况下，原文 f 的概率，因此称为翻译模型， $p(e)$ 用于描述译文句子的概率，称为语言模型。统计机器翻译的原理可由图 1-2 表示。

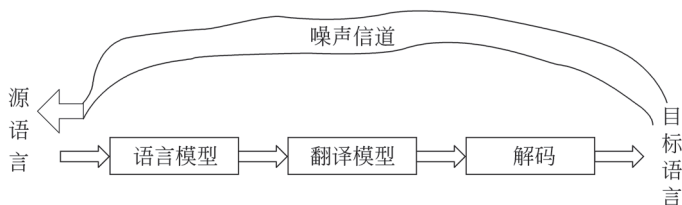


图 1-2 统计机器翻译的原理

根据模型中语言粒度不同，统计机器翻译的发展先后经历了词汇模型（word-based model），短语模型（phrase-based model）和句法模型（syntax-based model）等几种统计翻译模型。词汇模型的基本翻译单位是词汇，由词汇的翻译结果组织成句子，因此词汇模型的训练语料是词汇级对齐的翻译句对。而要得到大规模词对齐的翻译语料的代价较高。IBM 曾提出了四种对齐模型，可由算法实现自动对齐，并基于对齐模型形成了一些开源的词对齐程序，如 GIZA++ 等。但是，因为翻译不是简单的词到词的对译，基于词对齐的翻译与人工翻译的理念不符，所以词对齐模型的翻译质量并不很高。短语模型是在此基础上进行的改进，基

本翻译单位是短语。训练短语翻译模型的语料自然要求是短语对齐的双语数据。后期提出的句法模型则首先需要进一步对翻译句子进行句法分析，然后再从语法层次进行对齐和翻译。随着基本翻译单位的变化，统计翻译质量得到进一步提升，同时模型的复杂度也大大增加了。

1.2.4 神经网络机器翻译

神经网络机器翻译是基于大规模语料训练一个复杂的神经网络结构来将输入的源语言句子翻译成目标语句子。神经网络机器翻译（Neural Machine Translation, NMT）的基本翻译单位是句子，即输入源语言句子后直接得到目标语句子，也称为端到端（end-to-end）的翻译方法。神经网络机器翻译的突出优点是能够从大量翻译语料中自动完成翻译特征的学习，并保存为网络节点之间的连接权重。尤其近两年来，自然语言处理的序列模型中增加了注意力机制（attention mechanism），使得机器能够在不同网络层关注到语义上不同层面的东西。图 1-3 是 2016 年谷歌的神经机器翻译 GNMT 模型。GNMT 系统主要由编码（encoder）和译码（decoder）两部分构成。图 1-3 左侧为编码部分，右侧为译码部分。

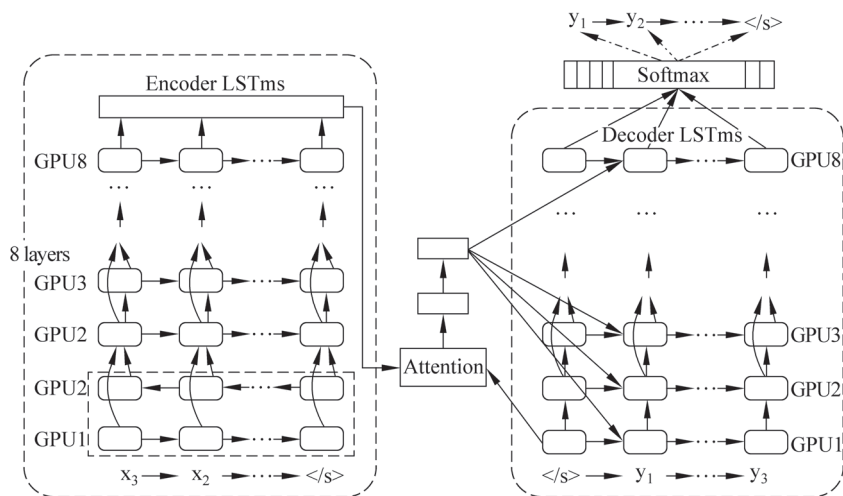


图 1-3 2016 年谷歌的神经网络机器翻译系统框架

2017 年提出的 Transformer 模型（Vaswani *et al.*, 2017）是一个以

注意力机制为核心的翻译模型，大大提高了翻译的性能。以此为基础，2018年又结合预训练（pretraining）提出了BERT模型（Devlin *et al.*, 2018），以及后来的ALBERT模型（Lan *et al.*, 2019）。这些模型都是有着百万数量级参数的神经网络架构，对计算机的算力提出了更高的要求。图 1-4 是 Transformer 模型的架构图。

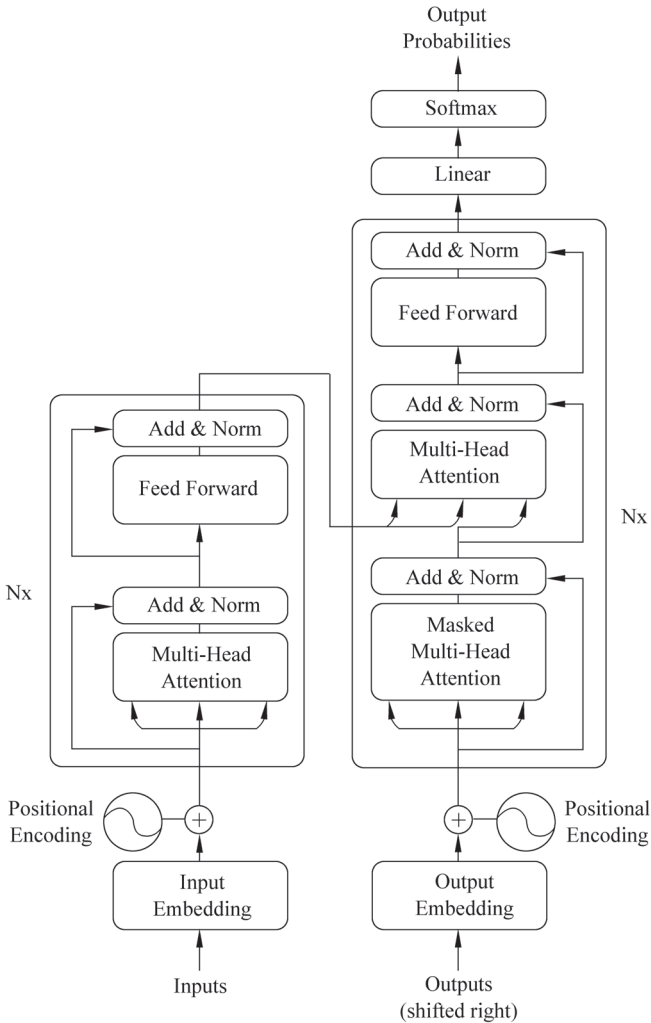


图 1-4 Transformer 模型的架构

神经网络机器翻译的卓越表现依靠三点：一是超大规模的训练数据，二是深度网络模型，三是超强的算力（秦颖，2018）。训练语料数据的规模一般要在千万词以上，网络的层数也在不断提高，有几十层到上百层不等。在如此大量数据集上训练如此复杂的网络结构，普通计算机是无法完成的，因此训练神经网络机器翻译模型需要的硬件条件也大幅提高了。

尽管神经网络机器翻译质量取得了较显著的提升，但是翻译结果的可解释性不如前几种翻译方法，人们无法说明这些网络连接权重的含义和对应翻译结果之间的必然联系。对神经网络翻译的研究出现了很多子方向，例如从网络模型的结构到解码方法、从字符翻译到和先验知识的融合、从预训练到结果的可视化和可解释性、从多语种翻译到翻译的公平性和多样性，都有很多研究成果。

神经网络虽然被实践证明是强大的，但是从理论角度对深度学习网络原理的研究还处在起步阶段，对于神经网络的结构形式如何影响其功能，我们还知之甚少。研究者虽可以采取实验的方式对复杂的网络参数获得经验值，却不能从理论上解释为什么这么设置。尽管 Rolnick & Tegmark（2018）的研究初步说明了网络的深度和神经元个数的关系，但大多数网络的深度和宽度参数以及神经元之间的连接方式、连接权重的设置等都是盲目的。相比之下，传统机器翻译的问题一般可分为两个方面：一方面来自对源语言分析错误，比如介词短语附着（pp-attachment）关系的分析错误会导致译文的介词短语翻译错误；另一方面来自对目标语的语言组织问题，比如语序和搭配词的选择不当会造成译文不流畅。然而，现在神经网络机器翻译并不对源语言的句子进行语言分析，翻译是“黑盒”的过程，既不能解释结果正确的原因，也不能解释错误的原因。因此，神经网络机器翻译的不可解释性对发现并评估翻译错误提出了挑战。

词嵌入（word embedding）增强了词汇表示语义的能力，预训练进一步提高了词向量在语境中表达语义的能力。由于有限数量的上下文仍不能体现词汇的全部含义和约束条件，因此多义词的翻译仍未完全得到解决，完整语义翻译的准确性还相对有限。另外，低频词、少见词和开放性词汇的翻译问题仍然是机器翻译的难题，神经网络机器翻译尚有不少问题没有得到很好的解决。

1.3 机器翻译整体特征

纵观机器翻译的原理，无论是基于规则的方法，还是当前复杂的神经网络的方法，都是基于实际语料和计算模型进行的，也就是从经验中通过计算得到的结果。计算机从根本上看并没有在理解自然语言的进程中取得重大突破，所有机器翻译都不是基于理解的翻译，这一点和人工翻译有根本的不同，机器不能像人一样先对源语言进行分析理解后进行翻译。人工翻译通常是自上而下的翻译，先从整体上理解原文，分析语域和语境的具体情况，再到句子一级的翻译，因而可以做到语义层面和语言功能层面的等价。机器只是从语言表层处理语言，难以转换源语言深层要表达的内容。对于是否一定要先全面深入理解语言才能翻译，当然也有很多人持有不同的观点。Ben-Ari *et al.* (1988) 和 Isabelle & Bourbeau (1985) 就认为，人工翻译其实并不一定要全面理解文字，很多分歧可以保存在译文中或者通过译文旁路 (bypass) 过去，留给读者去理解。

随着单语和双语语料数据规模的增多和学习模型复杂度的不断提升，机器翻译未来能达到的极限高度还是未解之谜。机器翻译的质量不断提升，译文越来越流利和准确，这一发展趋势也是毋庸置疑的。