

机器翻译

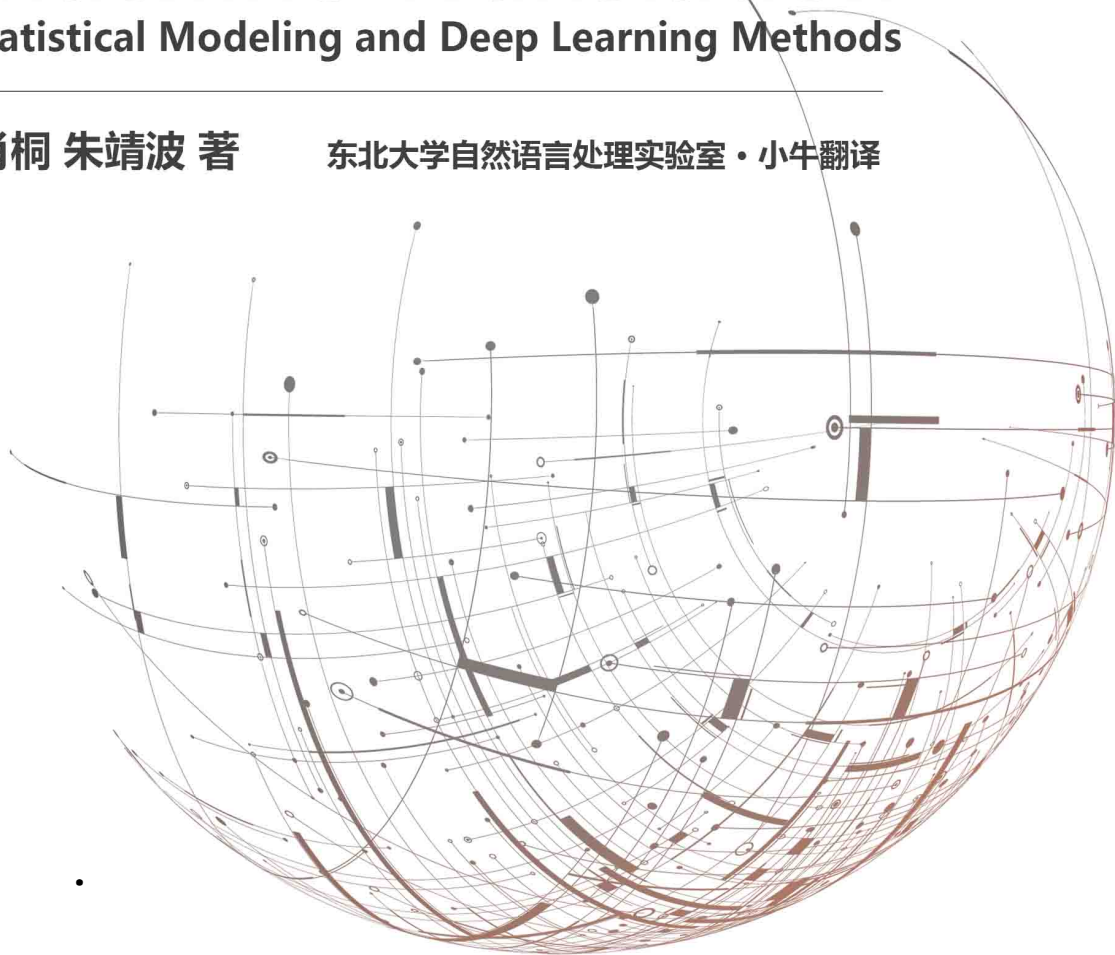
统计建模与深度学习方法

Machine Translation

Statistical Modeling and Deep Learning Methods

肖桐 朱靖波 著

东北大学自然语言处理实验室 · 小牛翻译



Copyright © 2020 肖桐 朱靖波

东北大学自然语言处理实验室 · 小牛翻译

[HTTPS://OPENSOURCE.NIUTRANS.COM/MTBOOK/INDEX.HTML](https://opensource.niutrans.com/mtbook/index.html)

[HTTPS://GITHUB.COM/NIUTRANS/MTBOOK](https://github.com/niuTrans/MTBook)

Licensed under the Creative Commons Attribution-NonCommercial 4.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/4.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

May 15, 2020

在此感谢所有为本书做出贡献的人

曹润柘、曾信、孟霞、单韦乔、姜雨帆、王子扬、刘辉、许诺、李北、刘继强、张哲
旸、周书含、周涛、张裕浩、李炎洋、林野、陈贺轩、刘晓倩、牛蕊、田丰宁、杜权

导 读

让计算机进行自然语言的翻译是人类长期的梦想，也是人工智能的终极目标之一。自上世纪九十年代起，机器翻译迈入了基于统计建模的时代，发展到今天，深度学习等机器学习方法已经在机器翻译中得到了大量的应用，取得了令人瞩目的进步。

在这个时代背景下，对机器翻译的模型、方法和实现技术进行深入了解是自然语言处理领域研究者和实践者所渴望的。本书全面回顾了近三十年内机器翻译的技术发展历程，并围绕统计建模和深度学习两个主题对机器翻译的技术方法进行了全面介绍。在写作中，笔者力求用朴实的语言和简洁的实例阐述机器翻译的基本模型和方法，同时对相关的技术前沿进行讨论。本书可以供计算机相关专业高年级本科生及研究生学习之用，也可以作为自然语言处理，特别是机器翻译领域相关研究人员的参考资料。

本书共分为七个章节，章节的顺序参考了机器翻译技术发展的时间脉络，同时兼顾了机器翻译知识体系的内在逻辑。各章节的主要内容包括：

- 第一章：机器翻译简介
- 第二章：词法、语法及统计建模基础
- 第三章：基于词的机器翻译模型
- 第四章：基于短语和句法的机器翻译模型
- 第五章：人工神经网络和神经语言建模
- 第六章：神经机器翻译模型
- 第七章：神经机器翻译实战 —— 参加一次比赛

其中，第一章是对机器翻译的整体介绍。第二章和第五章是对统计建模和深度学习方法的介绍，分别建立了两个机器翻译范式的基础知识体系——统计机器翻译和神经机器翻译。统计机器翻译部分（第三、四章）涉及早期的基于单词的翻译模型，以及本世纪初流行的基于短语和句法的翻译模型。神经机器翻译（第六、七章）代表了当今机器翻译的前沿，内容主要涉及了基于端到端表示学习的机器翻译建模方法。特别的，第七章对一些最新的神经机器翻译方法进行了讨论，为相关科学问题的研究和实用系统的开发提供了可落地的思路。图1展示了本书各个章节及核心概念之间的关系。

用最简单的方式阐述机器翻译的基本思想是笔者所期望达到的目标。但是，书中不可避免会使用一些形式化定义和算法的抽象描述，因此，笔者尽所能通过图例进行解释（本书共 320 张插图）。不过，本书所包含的内容较为广泛，难免会有疏漏，望读者海涵，并指出不当之处。

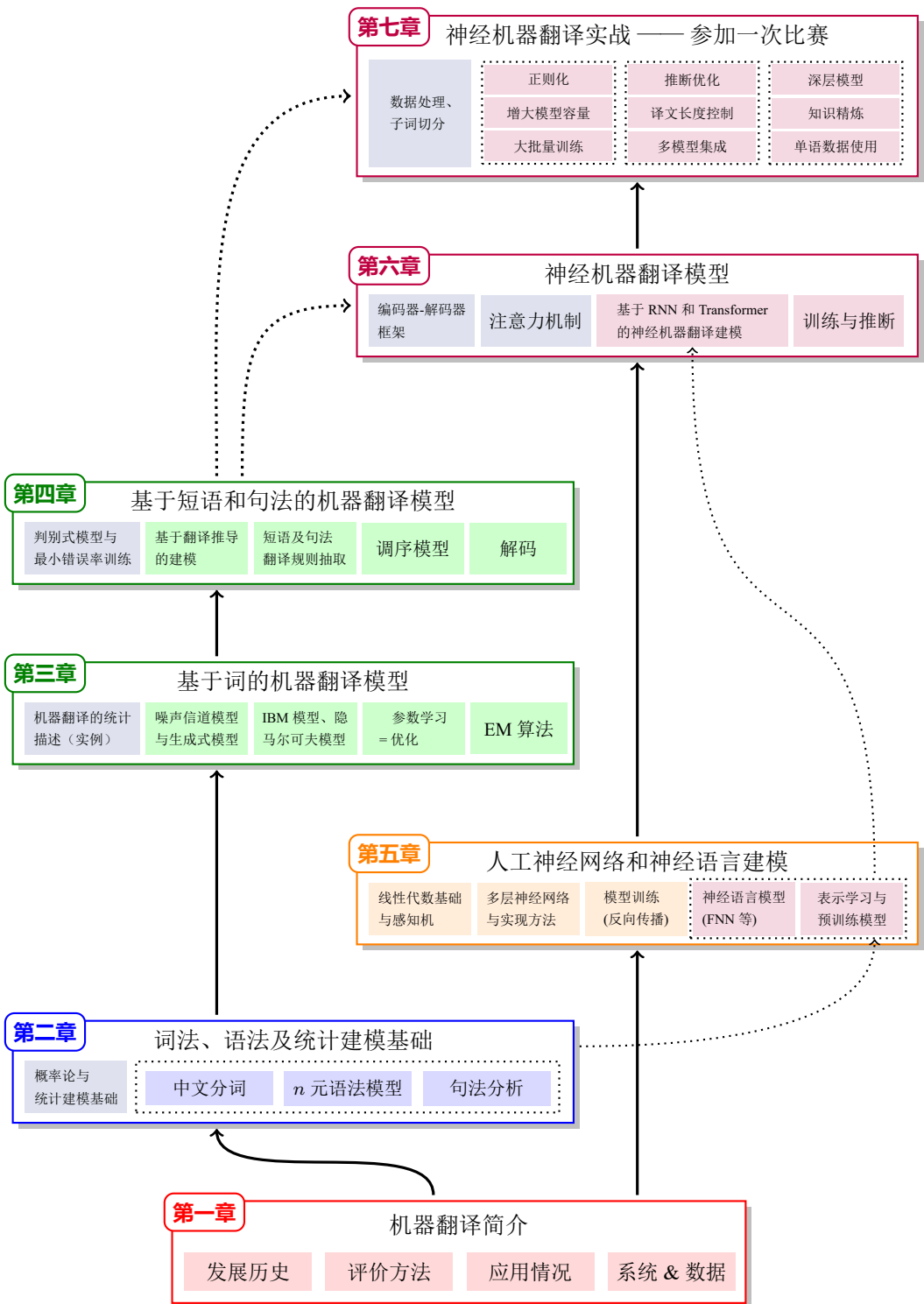


图 1: 本书各章节及核心概念关系图



Contents

I

机器翻译基础

1	机器翻译简介	11
1.1	机器翻译的概念	11
1.2	机器翻译简史	14
1.2.1	人工翻译	14
1.2.2	机器翻译的萌芽	15
1.2.3	机器翻译的受挫	16
1.2.4	机器翻译的快速成长	17
1.2.5	机器翻译的爆发	18
1.3	机器翻译现状	19
1.4	机器翻译方法	21
1.4.1	基于规则的机器翻译	21
1.4.2	基于实例的机器翻译	22
1.4.3	统计机器翻译	23
1.4.4	神经机器翻译	24
1.4.5	对比分析	25

- 1.5 翻译质量评价 26**
 - 1.5.1 人工评价 26
 - 1.5.2 自动评价 27
- 1.6 机器翻译应用 30**
- 1.7 开源项目与评测 32**
 - 1.7.1 开源机器翻译系统 32
 - 1.7.2 常用数据集及公开评测任务 36
- 1.8 推荐学习资源 38**

机器翻译基础

1	机器翻译简介	11
1.1	机器翻译的概念	
1.2	机器翻译简史	
1.3	机器翻译现状	
1.4	机器翻译方法	
1.5	翻译质量评价	
1.6	机器翻译应用	
1.7	开源项目与评测	
1.8	推荐学习资源	

1. 机器翻译简介

1.1 机器翻译的概念

从广义上来讲，“翻译”是指把一个事物转化为另一个事物的过程。这个概念多使用在对序列的转化上，比如，计算机程序的编译、自然语言文字翻译、蛋白质生物合成等。在程序编译中，高级语言编写的程序经过一系列的处理后转化为可执行的目标程序，这是一种从高级程序语言到低级程序语言的“翻译”。在人类语言的翻译中，一种语言文字通过人脑转化为另一种语言表达，这是一种自然语言的“翻译”。在蛋白质合成的第一步，RNA 分子序列转化为特定的氨基酸序列，这是一种生物学遗传信息的“翻译”。甚至说给上联对出下联、给一幅图片写出图片的主题等都可以被看作是“翻译”的过程。

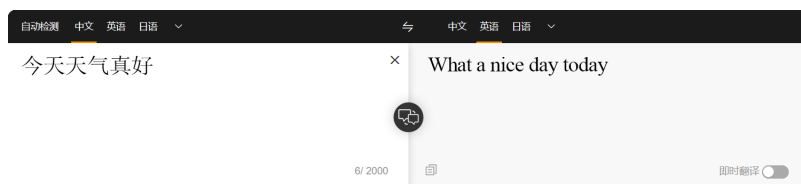


图 1.1: 通过计算机将中文翻译为英文

这里更加关注人类语言之间的翻译问题，即自然语言的翻译。如图1.1所示，通过计算机可以将一段中文文字自动转化为英文文字，中文被称为**源语言**（Source Language），英文被称为**目标语言**（Target Language）。

一直以来，文字的翻译往往是由人工完成。让计算机像人一样进行翻译似乎还是电影中的桥段，因为人们很难想象语言的多样性和复杂性可以用计算机语言进行描述。但是时至今日，人工智能技术的发展已经大大超越了人类传统的认知，用计算机进行自动翻译也不再是一种梦想，它已经深入到人们生活的很多方面，并发挥着重要作用。而这种由计算机进行自动翻译的过程也被称作**机器翻译**（Machine Translation）。类似地，自动翻译、智能翻译、多语言自动转换等概念也是指同样的事情。如果将今天的机器翻译和人工翻译进行对比，可以发现机器翻译系统所生成的译文还并不完美，甚至有时翻译质量非常差，但是它的生成速度快且成本低廉，更为重要的是机器翻译系统可以从大量数据中不断学习和进化。

人工翻译尽管精度很高，但是费时费力。当需要翻译大量的文本且精度要求不那么高时，比如海量数据的浏览型任务，机器翻译的优势就体现出来了。对于人工作业无法完成的事情，使用机器翻译可能只需花费几个小时甚至几分钟就能完成。这就类似于拿着锄头耕地种庄稼和使用现代化机器作业之间的区别。

实现机器翻译往往需要多个学科知识的融合，如数学、语言学、计算机科学、心理学等等。而最终呈现给使用者的是一套软件系统——机器翻译系统。通俗来讲，机器翻译系统就是一个可以在计算机上运行的软件工具，与人们使用的其他软件一样，只不过机器翻译系统是由“不可见的程序”组成。虽然这个系统非常复杂，但是呈现出来的展示形式却很简单，比如输入是待翻译的句子或文本，输出是译文句子或文本。

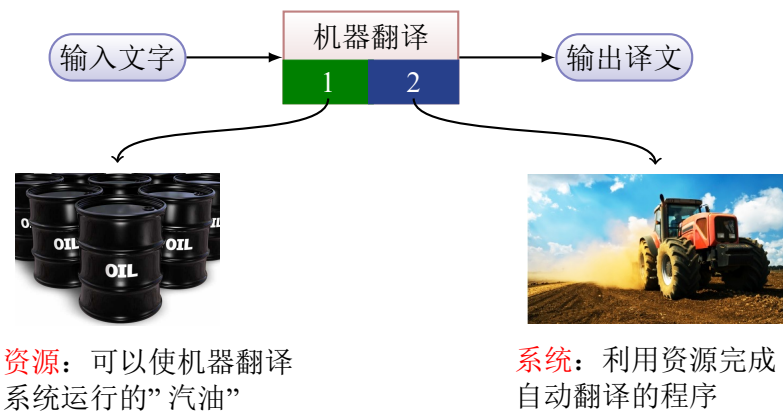


图 1.2: 机器翻译系统的组成

用机器进行翻译的想法可以追溯到电子计算机产生之前，发展过程中也经历了多个范式的变迁，现代机器翻译系统大多是基于数据驱动的方法——从数据中自动学习翻译知识，并运用这些知识对新的文本进行翻译。如图1.2所示，机器翻译系统通常由两部分组成：

- **资源**：如果把机器翻译系统比作一辆汽车，资源就好比是可以使汽车运行的“汽油”，它包括很多内容，如翻译规则、双（单）语数据、知识库等翻译知识，且

这些“知识”都是计算机可读的。值得一提的是,如果没有翻译资源的支持,任何机器翻译系统都无法运行起来。

- **系统:** 机器翻译算法的程序实现被称作系统,也就是机器翻译研究人员开发的软件。无论是翻译规则、翻译模板还是统计模型中的参数都需要通过机器翻译系统进行读取和使用。

构建一个强大的机器翻译系统需要“资源”和“系统”两方面共同作用。在资源方面,随着语料库语言学的发展,已经有大量高质量的双语和单语数据(称为语料)被整理并且电子化存储,研发机器翻译系统所需要的语料基础已经具备。特别是像英语、汉语等世界主流语种,相关语料资源已经非常丰富,这也大大加速了相关研究的进展。当然,对于一些稀缺资源语种或者特殊的领域,语料库仍然匮乏,但是这些并不影响机器翻译领域整体的发展速度。在现有语料库的基础上,很多研究者可以把精力集中在“系统”上。但是,机器翻译并非易事,有以下几方面挑战:

- **自然语言翻译问题的复杂性极高。**语言是人类进化的最高成就之一,自然语言具有高度的概括性、灵活性、多样性,这些都很难用几个简单的模型和算法进行描述。因此,翻译问题的数学建模和计算机程序实现难度很大。虽然近几年 Alpha Go 等人工智能系统在围棋等领域取得了令人瞩目的成绩,但是,相比翻译来说,围棋等棋类任务仍然“简单”,比如,对于一个句子,其潜在的译文几乎是不可穷尽的,即使同一句话不同人的理解也不尽相同,甚至在翻译一个句子、一个单词的时候,要考虑整个篇章的上下文语境,这些问题都不是传统棋类任务所具备的。
- **计算机的“理解”与人类的“理解”存在鸿沟。**人类一直希望把自己翻译时所使用的知识描述出来,并用计算机程序进行实现,例如早期基于规则的机器翻译方法就源自这个思想。但是,经过实践发现,人和计算机在“理解”自然语言上存在着明显差异。首先,人类的语言能力是经过长时间在多种外部环境因素共同作用下形成的,这种能力很难直接准确地表达。况且人类的语言知识本身就很难描述,更不用说让计算机来理解;其次,人和机器翻译系统理解语言的目的不一样。人理解和使用语言是为了进行生活和工作,而机器翻译系统更多的是为了对某些数学上定义的目标函数进行优化。也就是说,机器翻译系统关注的是翻译这个单一目标,而并不是像人一样进行复杂的活动;此外,人和计算机的运行方式有着本质区别。人类语言能力的生物学机理与机器翻译系统所使用的计算模型本质上是不同的,机器翻译系统使用的是其自身能够理解的“知识”,比如,统计学上的词语表示。这种“知识”并不需要人来理解,当然从系统开发的角度,计算机也并不需要理解人是如何思考的。
- **单一的方法无法解决多样的翻译问题。**首先,语种的多样性会导致任意两种语言之间的翻译实际上都是不同的翻译任务。比如,世界上存在的语言多达几千种,如果选择任意两种语言进行互译就产生上百万种翻译方向。虽然已经有研究者

尝试用同一个框架甚至同一个翻译系统进行全语种的翻译，但是这类系统离真正可用还有很远的距离；其次，不同的领域，不同的应用场景对翻译也有不同的需求。比如，文学作品的翻译和新闻的翻译就有不同、口译和笔译也有不同，类似的情况不胜枚举。机器翻译要适用于多样的需求，这些又进一步增加了计算机建模的难度；再次，对于机器翻译来说，充足的高质量数据是必要的，但是不同语种、不同领域、不同应用场景所拥有的数据量有明显差异，甚至很多语种几乎没有可用的数据，这时开发机器翻译系统的难度可想而知。值得注意的是，现在的机器翻译还无法像人类一样在学习少量样例的情况下进行举一反三，因此数据稀缺情况下的机器翻译也给研究者带来了很大的挑战。

显然，实现机器翻译并不简单，甚至有人把机器翻译看作是实现人工智能的终极目标。幸运的是，今天的机器翻译无论从技术方法上还是从应用上都有了巨大的飞跃，很多问题在不断被求解。如果你看到过十年前机器翻译的结果，再对比今天的结果，一定会感叹翻译质量的今非昔比，很多译文已经非常准确且流畅。从当今机器翻译的前沿技术看，近三十年机器翻译的进步更多的得益于基于数据驱动方法和统计建模方法的使用。特别是近些年深度学习等基于表示学习的端到端方法使得机器翻译的水平达到了新高度。因此，本书将会对基于统计建模和深度学习方法的机器翻译模型、方法和系统实现进行全面介绍和分析，希望这些内容可以对相关内容的学习和科研工作提供参考。

1.2 机器翻译简史

虽然翻译这个概念在人类历史中已经存在了上千年，但机器翻译发展至今只有七十余年的历史。纵观机器翻译的发展，历程曲折又耐人寻味，可以说了解机器翻译的历史对我们深入理解相关技术方法会有很好的启发，甚至对我们了解整个自然语言处理领域的发展也有启示作用。

1.2.1 人工翻译

人类形成语言文字的过程中逐渐形成了翻译的概念。一个著名的标志性证据是罗塞塔石碑（Rosetta Stone），如图1.3所示。这个石碑制作于公元前 196 年，据说是可供考证的最久远的记载平行文字的历史遗迹。石碑由上至下刻有同一段埃及国王诏书的三种语言版本，最上面是古埃及象形文，中间是埃及草书，最下面是古希腊文。可以明显看出石碑上中下雕刻的文字的纹理是不同的。尽管用不同的语言文字描述同一件事在今天看来很常见，但是这在生产力低下的两千年前是很罕见的。很多人认为罗塞塔石碑是标志翻译或人工翻译的一个起点。目前罗塞塔石碑保存于大英博物馆，并成为该馆最具代表性的镇馆之宝之一。

随后，更多的翻译工作在文化和知识传播中开展。其中一个典型代表是宗教文献的翻译。在人类的历史长河中，宗教是人类意识形态的一个重要载体。为了宣传

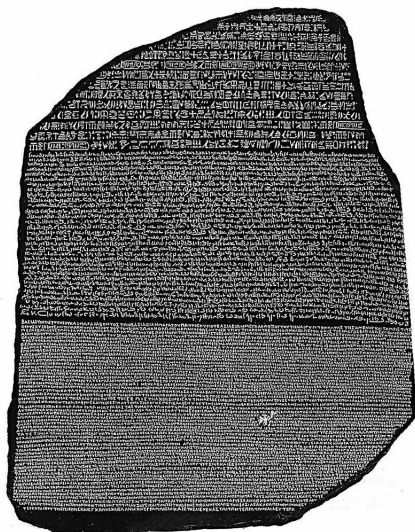


图 1.3: 罗塞塔石碑

教义，人们编写了大量的宗教文献。在西方，一项最早被记录的翻译活动是将旧约圣经（希伯来文及埃兰文）翻译为希腊文版本。迄今为止人类历史上翻译版本最多的书就是圣经。在中国唐代，有一位世界性的重量级文化人物——玄奘，他不仅是佛学家、旅行家，还是翻译家。玄奘西行求法归来后把全部的心血和智慧奉献给了译经事业，在助手们的帮助下，共翻译佛教经论 74 部，1335 卷，每卷万字左右，合计 1335 万字，占去整个唐代译经总数的一半以上，树立了我国古代翻译思想的光辉典范。

翻译在人类历史长河中起到了重要的作用。一方面，由于语言文字、文化和地理位置的差异性，使得翻译成为一个重要的需求；另一方面，翻译也加速了不同文明的融会贯通，促进了世界的发展。今天，翻译已经成为重要的行业之一，包括各个高校也都设立了翻译及相关专业，相关人才不断涌现。据《2019 年中国语言服务行业发展报告》统计：全球语言服务产值预计将首次接近 500 亿美元；中国涉及语言服务的在营企业 360,000 余家，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；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（MTI）和翻译本科（BTI）专业的院校分别有 250 余所和 280 余所，MTI 累计招生数达 6 万余人 [40]。当然，面对着巨大的需求，如何使用技术手段提高人工翻译效率，比如：机器辅助翻译，也是人工翻译和机器翻译领域需要共同探索的方向。

1.2.2 机器翻译的萌芽

人工翻译已经存在了上千年，而机器翻译又起源于什么时候呢？机器翻译的兴起与发展可以说充满了跌宕起伏，整个发展史可以分为萌芽期、受挫期、快速成长期和爆发期四个阶段。

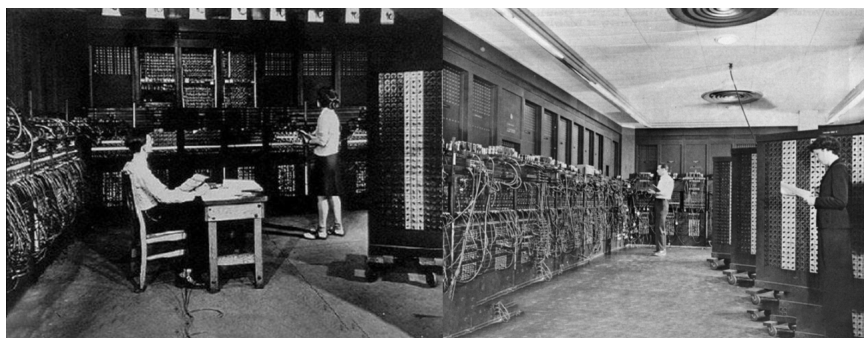


图 1.4: 世界上第一台通用电子数字计算机“埃尼阿克”(ENIAC)

世界上第一台通用电子数字计算机在 1946 年研制成功(图1.4)。但在上世纪 30 年代使用计算模型进行自动翻译的思想就开始萌芽,当时法国科学家 G.B. Artsouni 提出了用机器来进行翻译的想法。

第二次世界大战使得数学和密码学相当发达,由于战争的需要,在那个时代消息传递变的更为隐秘,对文字进行加密和解密成为重要的军事需求。因此,有人提出是否能用密码学的技术或方法解决人类语言的翻译,比如把汉语看成英语的一个加密文本,汉语翻译成英语就类似于解密的过程。当然这只是最初的想法。第一次提出机器翻译这个概念是在 1949 年,当时 W. Weaver 撰写了一篇名为《翻译》的备忘录,正式开创了机器翻译(Machine Translation)的概念,这个概念一直沿用至今。当然,在那个年代进行机器翻译研究还有很多条件不具备,包括使用加密解密技术进行自动翻译的很多尝试很快也被验证是不可行的。不过,这些早期的探索为后来机器翻译的发展提供了思想的火种。

1.2.3 机器翻译的受挫

随着电子计算机的发展,研究者开始尝试使用计算机来进行自动的翻译。但是事情并不总是一帆风顺,怀疑论者对机器翻译一直存有质疑,并很容易找出一些机器翻译无法解决的问题。自然地,人们也期望能够客观地评估一下机器翻译的可行性。当时美国基金资助组织委任自动语言处理咨询会承担了这项任务。

经过近两年的调查与分析,该委员会于 1966 年 11 月公布了一个题为《语言与机器》的报告(图1.5),简称 ALPAC 报告。该报告全面否定了机器翻译的可行性,为机器翻译的研究泼了一盆冷水。

随后美国政府终止了对机器翻译研究的支持,这导致整个产业界和学术界对机器翻译都开始回避。大家觉得机器翻译像伪科学,无论是发表论文还是申请项目都很难得到支持。没有了政府的支持,企业也无法进行大规模投入,机器翻译的研究就此受挫。

从历史上看,包括机器翻译在内很多人工智能领域在那个年代并不受“待见”,其主要原因在于当时的技术水平还比较低,而大家又对机器翻译等技术的期望过高。最

后发现，当时的机器翻译水平无法满足实际需要，因此转而排斥它。但是，也正是这一盆冷水，让人们可以更加冷静的思考机器翻译的发展方向，为后来的爆发蓄力。

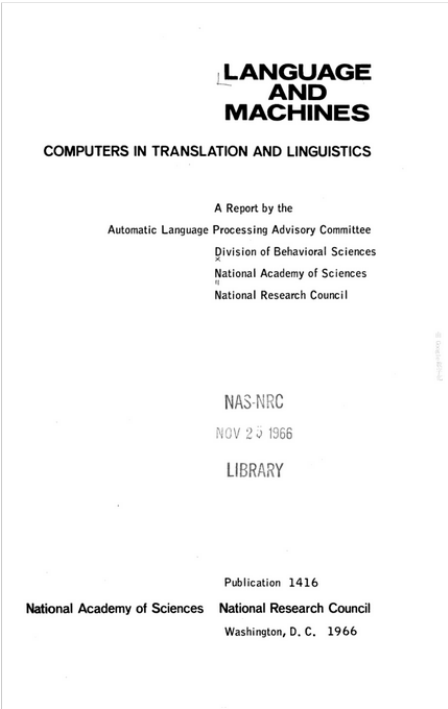


图 1.5: ALPAC 报告

1.2.4 机器翻译的快速成长

事物发展都是螺旋式上升的，机器翻译也是一样。上世纪 70 年代中后期，特别是 80 年代到 90 年代初，国家之间往来日益密切，而不同语言之间形成的交流障碍愈发严重，传统的人工作业方式已经远远不能满足需求。与此同时，语料库语言学的发展也为机器翻译提供了新的思路。其中，随着传统纸质文字资料不断电子化，计算机可读的语料越来越多，这使得人们可以用计算机对语言规律进行统计分析。另一方面，随着可用数据越来越多，用数学模型描述这些数据中的规律并进行推理逐渐成为可能。这也衍生出一类数学建模方法——**数据驱动**（Data-Driven）的方法。这类方法也成为了随后出现的统计机器翻译的基础。

传统的机器翻译方法，都需要人来书写规则，虽然对少部分句子具有较高的翻译精度，但这类方法对翻译现象的覆盖度有限，而且对规则或者模板中的噪声非常敏感，系统健壮性差。而基于数据驱动的方法不依赖于人写的规则，机器翻译的建模、训练和推断都可以自动地从数据中学习。这使得整个机器翻译的范式发生了翻天覆地的变化，比如，基于实例的方法和统计机器翻译就是在此期间兴起的。此外，这样的方法使得机器翻译系统的开发代价大大地降低。从上世纪 90 年代到本世纪初，统计机器翻译发展迅猛，很快成为了当时机器翻译研究与应用的代表性方法。一个标

志性的事件是谷歌推出了一个在线的免费自动翻译服务，也就是大家熟知的谷歌翻译。这使得机器翻译这种“高大上”的技术快速进入人们的生活，而不再是束之高阁的科研想法。随着机器翻译不断走向实用，机器翻译的应用也越来越多，这反过来进一步促进了机器翻译的研究进程。比如，在 2005-2015 年间，统计机器翻译这个主题几乎统治了 ACL 等自然语言处理相关方向顶级会议的论文，可见其在当时的影响力。

1.2.5 机器翻译的爆发

2005 年以后迎来了统计机器翻译发展的十年黄金时期，各种基于统计机器翻译模型层出不穷，经典的基于短语的模型和基于句法的模型也先后被提出。但是在 2013 年以后，机器学习的进步带来了机器翻译技术的进一步提升。特别是基于神经网络的深度学习方法在机器视觉、语音识别中被成功应用，带来性能的飞跃式提升。很快，相关模型和方法也被用于机器翻译。对于机器翻译来说，深度学习的成功也是一种必然，原因如下：

- 第一，端到端学习不依赖于过多的先验假设。在统计机器翻译时代，模型设计或多或少会对翻译的过程进行假设，称为隐藏结构假设。比如基于短语的模型假设：源语言和目标语言都会被切分成短语序列，这些短语之间存在某种对齐关系。这种假设既有优点也有缺点，一方面，假设可以有助于模型融入人类的先验知识，包括短语本身也借鉴了语言学相关的概念；另一方面，假设越多模型受到的限制也越多。如果假设是正确的，模型可以很好地描述问题。但如果假设错误，那么模型就可能产生偏差。深度学习不依赖于先验知识，也不需要手工设计特征，模型直接从输入和输出的映射上进行学习（端到端学习），这样也在一定程度上避免了隐藏结构假设造成的偏差。
- 第二，神经网络的连续空间模型有更强的表示能力。机器翻译中的一个基本问题是：如何表示一个句子？统计机器翻译把句子的生成过程看作是短语或者规则的推导，这本质上是一个离散空间上的符号系统。深度学习把传统的基于离散化的表示变成了连续空间的表示。比如，用实数空间的分布式表示代替了离散化的词语表示，而整个句子可以被描述为一个实数向量。这使得翻译问题可以在连续空间上描述，进而大大缓解了传统离散空间模型维度灾难等问题。更重要的是，连续空间模型可以用梯度下降等方法进行优化，具有很好的数学性质并且易于实现。
- 第三，深度网络学习算法的发展和 GPU（Graphics Processing Unit）等并行计算设备为训练神经网络提供了可能。早期的基于神经网络的方法一直没有在机器翻译甚至自然语言处理领域得到大规模应用，其中一个重要的原因是这类方法需要大量的浮点运算，而且以前计算机的计算能力无法达到这个要求。随着 GPU 等并行计算设备的进步，训练大规模神经网络也变为了可能。现在已经可

以在几亿、几十亿，甚至上百亿句对上训练机器翻译系统，系统研发的周期越来越短，进展日新月异。

今天，神经机器翻译已经成为新的范式，大有全面替代统计机器翻译之势。比如，从世界上著名的机器翻译比赛 WMT 和 CCMT 中就可以看出这个趋势。如图1.6所示，其中左图是 WMT19 全球机器翻译比赛的参赛队伍的截图，这些参赛队伍基本上都在使用深度学习完成机器翻译的建模。而在 WMT19 各个项目夺冠系统中（1.6右图），神经机器翻译也几乎一统天下。

值得一提的是，近些年神经机器翻译的快速发展也得益于产业界的关注。各大互联网企业和机器翻译技术研发机构都对神经机器翻译的模型和实践方法给予了很大贡献。比如，谷歌、微软、百度、搜狗、腾讯、阿里、有道、小牛翻译等企业凭借自身人才和基础设施方面的优势，先后推出了以神经机器翻译为内核的产品及服务，相关技术方法已经在大规模应用中得到验证，大大推动了机器翻译的产业化进程，而且这种趋势在不断加强，机器翻译的前景也更加宽广。

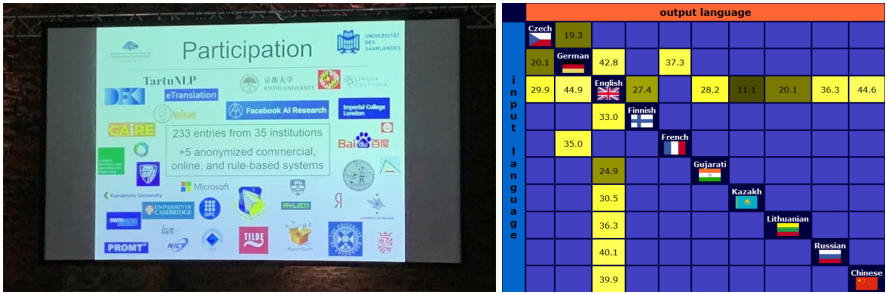


图 1.6: 国际机器翻译大赛 (左: WMT19 参赛队伍; 右: WMT19 最终个项目最好分数结果)

1.3 机器翻译现状

机器翻译技术发展到今天已经过无数次迭代，技术范式也经过若干次更替，近些年机器翻译的应用也如雨后春笋。但是大家都很好奇今天的机器翻译的质量究竟如何呢？乐观地说，在受限条件下，机器翻译的译文结果还是非常不错的，甚至可以接近人工翻译的结果。然而，在开放式翻译任务中，机器翻译的结果却并不理想。更严格来说，机器翻译的质量远没有达到人们所期望的完美的程度。对于有些人提到的“机器翻译代替人工翻译”也并不是事实。比如，在高精度同声传译任务中，机器翻译仍需要更多打磨；再比如，针对于小说的翻译，机器翻译还无法做到与人工翻译媲美；甚至有人尝试用机器翻译系统翻译中国古代诗词，这里更多的是娱乐的味道。但是毫无疑问的是，机器翻译可以帮助人类，甚至有朝一日可以代替一些低端的人工翻译工作。

图1.7展示了机器翻译和人工翻译质量的一个对比结果。在汉语到英语的新闻翻译任务中，如果对译文进行人工评价，五分制机器翻译的译文得 3.9 分，人工译文得 4.7 分（人的翻译也不是完美的）。可见，在这个任务中机器翻译表现不错，但是与

人还有一定差距。如果换一种方式评价，把人的译文作为参考答案，用机器翻译的译文与其进行比对，会发现机器翻译的得分只有 47 分（百分制）。当然，这个结果并不是说机器翻译的译文质量很差，它更多的是表明机器翻译系统可以生成一些与人工翻译不同的译文，机器翻译也具有一定的创造性。这也类似于，很多围棋选手都想向 AlphaGo 学习，因为智能围棋系统也可以走出一些人类从未走过的妙招。

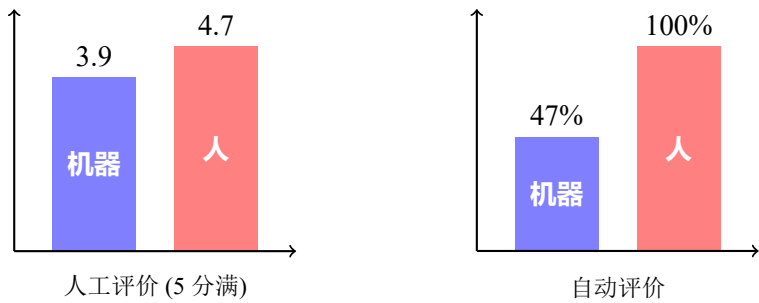


图 1.7: 机器翻译与人工翻译性能对比（汉英新闻领域翻译）

图1.8展示了一个真实的汉语到英语翻译实例。对比发现，机器翻译与人工翻译还是存在差距的，特别是在翻译一些具有感情色彩的词语时，机器翻译的译文缺一些味道。那么，机器翻译一点用都没有吗？显然不是。实际上，如果考虑翻译速度与翻译代价，机器翻译的价值是无可比拟的。还是同一个例子，翻译一篇短文如果人工翻译需要 30 分钟甚至更长时间，那么机器翻译仅仅需要两秒，换种情况思考，如果有 100 万篇这样的文档，其人工翻译的成本根本无法想象，消耗的时间更是难以计算，而计算机集群仅仅需要一天，而且只有电力的消耗。

源	文：从前有一个小岛，上面住着快乐、悲哀、知识和爱，还有其他各种情感。一天，情感们得知小岛快要下沉了。于是，大家都准备船只，离开小岛，只有爱决定留下来，她想坚持到最后一刻。过了几天，小岛真的要下沉了，爱想请人帮忙。
机器翻译：	Once upon a time there was an island <u>on which</u> lived happiness,sorrow,knowledge,love and other emotions. One day, <u>the</u> emotions learned that the island was going to sink.As a result,everyone pre- pared the boat and <u>left the island</u> . Only Love decided to stay.She <u>wanted to stick</u> to it until the last moment. After a few days, the island was really going to sink and love <u>wanted help</u> .
人工翻译：	Once upon a time, there was a small island <u>where</u> lived all kinds of emotions like JOY,SADNESS, KNOWLEDGE, and LOVE.One day, <u>these</u> emotions found that the island was sinking, so one by one they prepared the boat and <u>planned to leave</u> . None but LOVE chose to stay there. She <u>was deter- -mined to</u> persist till the last moment.A few days later, almost the whole island sunk into the sea, and LOVE had to <u>seek for help</u> .
源	文：这时，富裕乘着一艘大船经过。爱说：“富裕，你能带我走吗？”富裕答道：“不，我的船中有许多金银财宝，没有你的位置。”爱看见虚荣在一艘华丽的小船上，说：“虚荣，帮帮我吧！”“我帮不了你。你全身都湿透了，会弄坏我这漂亮的小船的。”悲哀来了，爱向他求助：“悲哀，让我跟你去吧！”“哦，……爱，我实在太悲哀了，想自己一个人呆一会儿！”悲哀答道。快乐走近爱的身边，但是她太快乐了，竟然没有听见爱在叫她！
机器翻译：	At this time, Richness <u>passed by</u> in a big ship. Love said, “Rich, can you take me away?” Richness replied, “No, <u>there are many treasures</u> of gold and silver in my ship,and there is no place for you.”

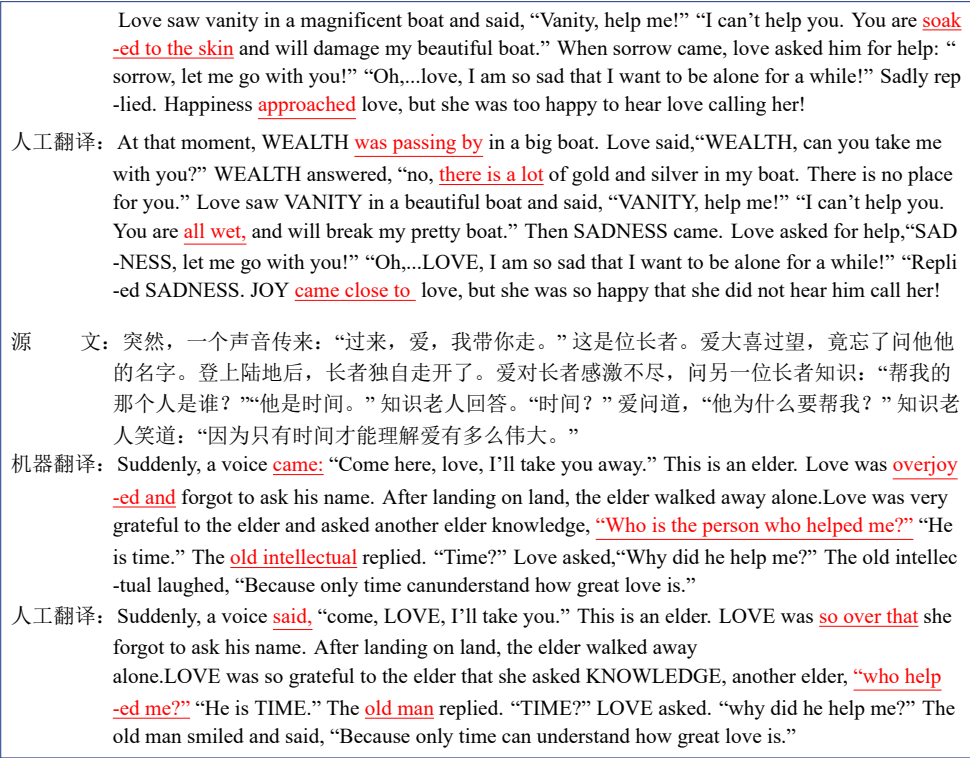


图 1.8: 机器翻译与人工翻译实例结果对比

1.4 机器翻译方法

为了对机器翻译技术有一个整体的认识, 这里对一些主要的机器翻译框架进行简要介绍。

1.4.1 基于规则的机器翻译

早期的机器翻译研究都是以基于规则的方法为主, 特别是在上世纪 70 年代, 以基于规则方法为代表的专家系统是人工智能中最具代表性的研究领域。它的主要思想是以词典和人工书写的规则库作为翻译知识, 用一系列规则的组合完成翻译。

图1.9展示了一个使用规则进行翻译的实例。这里, 利用一个简单的汉译英规则库完成对句子“我对你感到满意”的翻译。当翻译“我”时, 从规则库中找到规则 1, 该规则表示遇到单词“我”就翻译为“I”; 类似地, 也可以从规则库中找到规则 4, 该规则表示翻译调序, 即将单词“you”放到“be satisfied with”后面。可以看到, 这些规则的使用和进行翻译时所使用的思想非常类似, 可以说基于规则方法实际上在试图描述人类进行翻译的思维过程。

但是, 基于规则的机器翻译也存在问题。首先, 书写规则需要消耗大量人力, 规则库的维护代价极高; 其次, 规则很难涵盖所有的语言现象; 再有, 自然语言存在大



图 1.9: 基于规则的机器翻译的示例图（左：规则库；右：规则匹配结果）

量的歧义现象，规则之间也会存在冲突，这也导致规则数量不可能无限制增长。

1.4.2 基于实例的机器翻译

基于规则的方法更多地被使用在受限翻译场景中，比如受限词汇集的翻译。针对基于规则的方法存在的问题，基于实例的机器翻译于上世纪 80 年代中期被提出 [21]。该方法的基本思想是在双语句库中找到与待翻译句子相似的实例，之后对实例的译文进行修改，如替换、增加、删除等一系列操作，从而得到最终译文。这个过程可以类比人类学习并运用语言的过程：人会先学习一些翻译实例或者模板，当遇到新的句子时，会用以前的实例和模板作对比，之后得到新的句子的翻译结果。这也是一种举一反三的思想。



图 1.10: 基于实例的机器翻译的示例图（左：实例库；右：翻译结果）

图1.10展示了一个基于实例的机器翻译过程。它利用简单的翻译实例库与翻译词典完成对句子“我对你感到满意”的翻译。首先，使用待翻译句子的源语言端在翻译实例库中进行比较，根据相似度大小找到相似的实例“我对他感到高兴”。然后，标

记实例中不匹配的部分，即“你”和“他”，“满意”和“高兴”。再查询翻译词典得到词“你”和“满意”所对应的翻译结果“you”和“satisfied”，用这两个词分别替换实例中的“him”和“happy”，从而得到最终译文。

当然，基于实例的机器翻译也并不完美。首先，这种方法对翻译实例的精确度要求非常高，一个实例的错误可能会导致一个句型都无法翻译正确；其次，实例维护较为困难，实例库的构建通常需要单词级对齐的标注，而保证词对齐的质量是非常困难的工作，这也大大增加了实例库维护的难度；再次，尽管可以通过实例或者模板进行翻译，但是其覆盖度仍然有限。在实际应用中，很多句子无法找到可以匹配的实例或者模板。

1.4.3 统计机器翻译

统计机器翻译兴起于上世纪 90 年代 [2, 16] 它利用统计模型从单/双语语料中自动学习翻译知识。具体来说，可以使用单语语料学习语言模型，使用双语平行语料学习翻译模型，并使用这些统计模型完成对翻译过程的建模。整个过程不需要人工编写规则，也不需要从实例中构建翻译模板。无论是词还是短语，甚至是句法结构，统计机器翻译系统都可以自动学习。人更多的是定义翻译所需的特征和基本翻译单元的形式，而翻译知识都保存在模型的参数中。

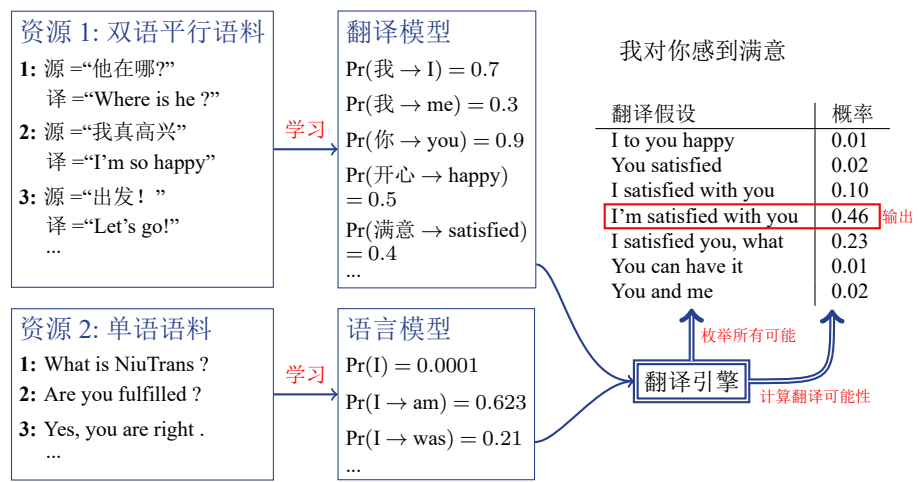


图 1.11: 统计机器翻译的示例图（左：语料资源；中：翻译模型与语言模型；右：翻译假设与翻译引擎）

图1.11展示了一个统计机器翻译系统运行的简单实例。整个系统需要两个模型：翻译模型和语言模型。其中，翻译模型从双语平行语料中学习翻译知识，得到短语表，其中包含各种词汇的翻译及其概率，这样可以度量源语言和目标语言片段之间互为翻译的可能性大小；语言模型从单语语料中学习目标语的词序列生成规律，来衡量目标语言译文的流畅性。最后，将这两种模型联合使用，翻译引擎来搜索尽可能多的翻译结果，并计算不同翻译结果的可能性大小，最后将概率最大的译文作为最终结果输出。这个过程并没有显性使用人工翻译规则和模板，译文的生成仅仅依

赖翻译模型和语言模型中的统计参数。

由于没有对翻译过程进行过多的限制，统计机器翻译有很灵活的译文生成方式，因此系统可以处理更加多样的句子。但是这种方法也带来了一些问题：首先，虽然并不需要人工定义翻译规则或模板，但统计机器翻译系统仍然需要人工定义翻译特征。提升翻译品质往往需要大量的特征工程，这导致人工特征设计的好坏会对系统产生决定性影响；其次，统计机器翻译的模块较多，系统研发比较复杂；再次，随着训练数据增多，统计机器翻译的模型（比如短语翻译表）会明显增大，在系统存储资源受限的情况下，这种模型不利于系统的正常使用。

1.4.4 神经机器翻译

随着机器学习技术的发展，基于深度学习的神经机器翻译逐渐兴起。自 2014 年开始，它在短短几年内已经在大部分任务上取得了明显的优势 [1, 29]。在神经机器翻译中，词串被表示成实数向量，即分布式向量表示。这样，翻译过程并不是在离散化的单词和短语上进行，而是在实数向量空间上计算，因此它对词序列表示的方式产生了本质的改变。通常，机器翻译可以被看作一个序列到另一个序列的转化。在神经机器翻译中，序列到序列的转化过程可以由**编码器-解码器**（encoder-decoder）框架实现。其中，编码器把源语言序列进行编码，并提取源语言中信息进行分布式表示，之后解码器再把这种信息转换为另一种语言的表达。

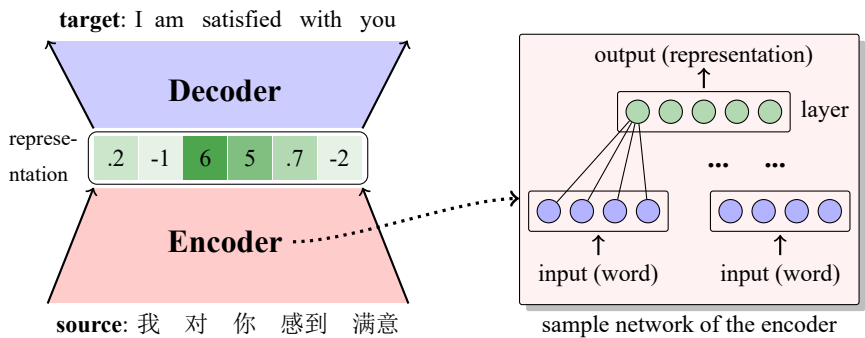


图 1.12: 神经机器翻译的示例图（左：编码器-解码器网络；右：编码器示例网络）

图1.12展示了一个神经机器翻译的实例。首先，通过编码器，源语言序列“我对你感到满意”经过多层神经网络编码生成一个向量表示，即图中的向量（0.2，-1，6，5，0.7，-2）。再将该向量作为输入送到解码器中，解码器把这个向量解码成目标语言序列。注意，目标语言序列的生成是逐词进行的（虽然图中展示的是解码器生成整个序列，但是在具体实现时是逐个单词生成目标语译文），产生某个词的时候依赖之前生成的目标语言的历史信息，直到产生句子结束符为止。

相比统计机器翻译，神经机器翻译的优势体现在其不需要特征工程，所有信息由神经网络自动从原始输入中提取。而且，相比离散化的表示，词和句子的分布式连续空间表示可以为建模提供更为丰富的信息，同时可以使用相对成熟的基于梯度

的方法优化模型。此外，神经网络的存储需求较小，天然适合小设备上的应用。但是，神经机器翻译也存在问题。首先，虽然脱离了特征工程，但神经网络的结构需要人工设计，即使设计好结构，系统的调优、超参数的设置等仍然依赖大量的实验；其次，神经机器翻译现在缺乏可解释性，其过程和人的认知差异很大，通过人的先验知识干预的程度差；再次，神经机器翻译对数据的依赖很大，数据规模、质量对性能都有很大影响，特别是在数据稀缺的情况下，充分训练神经网络具有挑战。

1.4.5 对比分析

不同机器翻译方法有不同的特点。表1.1对比了这些方法，不难看出：

- 规则系统需要人工书写规则并维护，人工代价较高。统计和神经网络方法仅需要设计特征或者神经网络结构，对人工依赖较少（语言相关的）。
- 基于实例、统计和神经网络的方法都需要依赖语料库（数据），其中统计和神经网络方法具有一定的抗噪能力，因此也更适合大规模数据情况下的机器翻译系统研发。
- 基于规则和基于实例的方法在受限场景下有较好的精度，但是在开放领域的翻译上统计和神经网络方法更具优势。

借用机器翻译著名学者董振东先生对不同机器翻译方法的评价¹：基于规则和实例的机器翻译是傻子（依赖一定人工，在匹配规则和模板的情况下翻译质量高，但是系统泛化能力有限），统计和神经机器翻译是疯子（只依赖数据，不依赖人工书写规则或者模板，系统健壮性强，但是精度不稳定且翻译过程难以人工干预）。

表 1.1: 不同机器翻译方法的对比

	规则	实例	统计	神经
人工写规则	是	否	否	否
人工代价	高	一般	几乎没有	几乎没有
数据驱动	否	是	是	是
依赖数据质量	N/A	高	低	较低
抗噪声能力	低	低	高	较高
使用范围	受限领域	受限领域	通用领域	通用领域
翻译精度	高	较高	不确定	不确定

从现在机器翻译的研究和应用情况来看，基于统计建模的方法（统计机器翻译和神经机器翻译）是主流。这主要是由于它们的系统研发周期短，通过搜集一定量的数据即可实现快速原型。但是随着互联网等信息的不断开放，低成本的数据获取可以让系统更快得以实现。特别是最近神经机器翻译凭借其高质量的译文，受到研究人员和开发者的广泛青睐。当然，对不同方法进行融合也是有价值的研究方向，也

¹这里遵循董振东先生的话进行了重新表述

有很多有趣的探索，比如无指导机器翻译中还是会同时使用统计机器翻译和神经机器翻译方法，这也是一种典型的融合多种方法的思路。

1.5 翻译质量评价

机器翻译质量的评价对于机器翻译的发展具有至关重要的意义。首先，评价的结果可以用于指导研究人员不断改进机器翻译结果，并找到最具潜力的技术发展方向。同时，一个权威的翻译质量评价指标可以帮助用户更有效地使用机器翻译的结果。

一般来说，机器翻译的翻译**质量评价**（Quality Evaluation）是指在参考答案或者评价标准已知的情况下对译文进行打分。这类方法可以被称作有参考答案的评价，包括人工打分、BLEU 等自动评价方法都是典型的有参考答案评价。相对的，**无参考答案的评价**（Quality Estimation）是指在没有人工评价和参考答案的情况下，对译文质量进行评估。这类方法可以被看作是对机器翻译译文进行质量“预测”，这样用户可以选择性的使用机器翻译结果。这里主要讨论有参考答案的评价，因为这类方法是机器翻译系统研发所使用的主要评价方法。

1.5.1 人工评价

顾名思义，人工评价是指专家根据翻译结果好坏对译文进行评价。一般会根据句子的准确度和流利度对其进行打分，这样能够准确评定出句子是否准确翻译出原文的意思以及句子是否通顺。在对一个句子进行评定时，一般由多个专家匿名打分后进行综合评定。人工评价是最能准确反映句子翻译质量的评价方式，但是其缺点也十分明显：需要耗费人力物力，而且评价的周期长，不能及时得到有效的反馈。因此在实际系统开发中，纯人工评价不会过于频繁的被使用，它往往和自动评价一起配合，帮助系统研发人员准确的了解当前系统的状态。

人工评价的策略非常多。考虑不同的因素，往往会使用不同的评价方案，比如：

- 打分：常用的方法是对每个译文进行百分制或者五分制打分，分数越高表明译文越好。更粗糙的方法有三分制，甚至两分制打分。注意，打分越细致，评价者的工作量越大，因此五分制或者三分制评价更适合快速获得评价结果。
- 是否呈现源语言文本：评价者在进行人工评价时可仅被提供源语言文本或参考译文、或同时提供源语言文本和参考译文。从评价的角度，参考译文已经能够帮助评价者进行正确评价，但是源语言文本可以提供更多信息帮助评估译文的准确性。
- 评价者选择：理想情况下，评价者应同时具有源语言和目标语言的语言能力。但是，很多场景下双语能力的评价者很难招募，因此这时会考虑使用目标语为母语的评价者。配合参考译文，单语评价者也可以准确地进行评价。

- 多个系统评价：如果有多个不同系统的译文需要评价，可以直接使用每个系统单独打分的方法。但是，如果仅仅是想了解不同译文之间的相对好坏，也可以采用竞评的方式，即对于每个句子，对不同系统根据译文质量进行排序，这样做的效率会高于直接打分，而且评价准确性也能够得到保证。
- 数据选择：评价数据一般需要根据目标任务进行采集，为了避免和系统训练数据重复，往往会搜集最新的数据。而且，评价数据的规模越大，评价结果越科学。常用的做法是搜集一定量的评价数据，之后从中采样出所需的数据。由于不同的采样会得到不同的评价集合，这样的方法可以复用多次，得到不同的测试集。
- 面向应用的评价：除了人工直接打分，一种更有效的方法是把机器翻译的译文嵌入到下游应用中，通过机器翻译对下游应用的改善效果评估机器翻译译文质量。比如，可以把机器翻译放入译后编辑流程中，通过对比翻译效率的提升来评价译文质量。再比如，把机器翻译放入线上应用中，通过点击率或者用户反馈来评价机器翻译的品质。

简而言之，研究者可以根据实际情况选择不同的人工评价方案，人工评价也没有统一的标准。WMT 和 CCMT 机器翻译评测都有配套的人工评价方案，可以作为业界的参考标准。

1.5.2 自动评价

由于人工评价费事费力，同时具有一定的主观性，甚至不同人在不同时刻面对同一篇文章的理解都会不同，因此自动评价也是机器翻译系统研发人员所青睐的方法。自动评价的方式虽然不如人工评价准确，但是具有速度快，成本低、一致性高的优点。而且随着评价技术的不断发展，自动评价方式已经具有了比较好的指导性，可以帮助使用者快速了解当前机器翻译译文的质量。在机器翻译领域，自动评价已经成为了一个重要的分支，提出的自动评价方法不下几十种。这里无法对这些方法一一列举，为了便于后续章节的描述，这里仅对具有代表性的一些方法进行简要介绍。

BLEU

目前使用最广泛的自动评价指标是 BLEU。BLEU 是 Bilingual Evaluation Understudy 的缩写，最早由 IBM 在 2002 年提出 [24]。通过采用 n -gram 匹配的方式评定机器翻译结果和参考译文之间的相似度，即机器翻译的结果越接近人工参考译文就认定它的质量越高。 n -gram 是指 n 个连续单词组成的单元，称为 n 元语法单元。 n 越大表示评价时考虑的匹配片段越大。

BLEU 的计算首先考虑待评价译文中 n -gram 在参考答案中的匹配率，称为 n -

gram 准确率 (n -gram Precision)。其计算方法如下：

$$P_n = \frac{\text{Count}_{\text{hit}}}{\text{Count}_{\text{output}}} \quad (1.1)$$

其中 $\text{Count}_{\text{hit}}$ 表示系统输出的译文中 n -gram 在参考答案中命中的次数, $\text{Count}_{\text{output}}$ 表示系统输出的译文中总共有多少 n -gram。为了避免同一个词被重复计算, BLEU 的定义中使用了截断的方式定义 $\text{Count}_{\text{hit}}$ 和 $\text{Count}_{\text{output}}$ 。例如：

■ **Example 1.1** Candidate: the the the the

Reference: The cat is standing on the ground ■

在引入截断方式之前, 该译文的 1-gram 准确率为 $4/4 = 1$, 这显然是不合理的。在引入截断的方式之后, “the” 在译文中出现 4 次, 在参考译文中出现 2 次, 截断操作则是取二者的最小值, 即 $\text{Count}_{\text{hit}} = 2$, $\text{Count}_{\text{output}} = 4$, 该译文的 1-gram 准确率为 $2/4$ 。

译文整体的准确率等于各 n -gram 的加权平均：

$$P_{\text{avg}} = \exp\left(\sum_{n=1}^N w_n \cdot \log P_n\right) \quad (1.2)$$

但是, 该方法更倾向于对短句子打出更高的分数。一个极端的例子是译文只有很少的几个词, 但是都命中答案, 准确率很高可显然不是好的译文。因此, BLEU 引入**短句惩罚因子** (Brevity Penalty, BP) 的概念, 对短句进行惩罚,

$$\text{BP} = \begin{cases} 1 & c > r \\ e^{(1-\frac{r}{c})} & c \leq r \end{cases} \quad (1.3)$$

其中 c 表示译文的句子长度, r 表示参考译文的句子长度。最终 BLEU 的计算公式为：

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{i=1}^N w_n \cdot \log P_n\right) \quad (1.4)$$

从机器翻译的发展来看, BLEU 的意义在于它给系统研发人员提供了一种简单、高效、可重复的自动评价手段, 在研发机器翻译系统时不需要依赖人工评价。同时, BLEU 也有很多创新之处, 包括引入 n -gram 的匹配, 截断计数和短句惩罚等等, 包括 NIST 等很多评价指标都是受到 BLEU 的启发。当然, BLEU 也并不完美, 甚至经常被人诟病。比如, 它需要依赖参考译文, 而且评价结果有时与人工评价不一致, 同时 BLEU 评价只是单纯的从匹配度的角度思考翻译质量的好坏, 并没有真正考虑

句子的语义是否翻译正确。但是，毫无疑问，BLEU 仍然是机器翻译中最常用的评价方法。在没有找到更好的替代方案之前，BLEU 还是机器翻译研究所使用的标准评价指标。

TER

TER 是 Translation Edit Rate 的缩写，是一种基于距离的评价方法，用来评定机器翻译结果的译后编辑的工作量 [28]。这里，距离被定义为将一个序列转换成另一个序列所需要的最少编辑操作次数。操作次数越多，距离越大，序列之间的相似性越低；相反距离越小，表示一个句子越容易改写成另一个句子，序列之间的相似性越高。TER 使用的编辑操作包括：增加，删除，替换和移位，其中增加，删除，替换操作计算得到的距离被称为编辑距离，并根据错误率的形式给出评分：

$$\text{score} = \frac{\text{edit}(c, r)}{l} \quad (1.5)$$

其中 $\text{edit}(c, r)$ 是指机器翻译生成的候选译文 c 和参考译文 r 之间的距离， l 是归一化因子，通常为参考译文的长度。在距离计算中所有的操作的代价都为 1。在计算距离时，优先考虑移位操作，再计算编辑距离，也就是增加、删除和替换操作的次数。直到增加、移位操作无法减少编辑距离时，将编辑距离和移位操作的次数累加得到 TER 计算的距离。例如：

■ **Example 1.2** Candidate: cat is standing in the ground

Reference: The cat is standing on the ground ■

将 Candidate 转换为 Reference，需要进行一次增加操作，在句首增加 “The”；一次替换操作，将 “in” 替换为 “on”。所以 $\text{edit}(c, r) = 2$ ，归一化因子 l 为 Reference 的长度 7，所以该参考译文的 TER 错误率为 $2/7$ 。

与 BLEU 不同，基于距离的评价方法是一种典型的“错误率”的度量，类似的思想也广泛应用于语音识别等领域。在机器翻译中，除了 TER 外，还有 WER, PER 等十分相似的方法，只是在“错误”的定义上略有不同。需要注意的是，很多时候，研究者并不会单独使用 BLEU 或者 TER，而是将两种方法融合，比如，使用 BLEU 与 TER 相减后的值作为评价指标。

基于检测点的评价

BLEU、TER 等评价指标可以对译文的整体质量进行评估，但是缺乏对具体问题的细致评价。很多时候，研究人员需要知道系统是否能够处理特定的问题，而不是得到一个笼统的评价结果。基于监测点的方法正是基于此想法 [27]。基于检测点的评价的优点在于对机器翻译系统给出一个总体评价的同时针对系统在各个具体问题上的翻译能力进行评估，方便比较不同翻译模型的性能。这种方法也被多次用于机器翻译比赛的质量评测。

基于检测点的评价是根据事先定义好的语言学检测点对译文的相应部分进行打分。如下是几个英中翻译中的检测点实例：

■ **Example 1.3** They got up at six this morning.

他们今天早晨六点钟起床。

检测点：时间词的顺序。 ■

■ **Example 1.4** There are nine cows on the farm.

农场里有九头牛。

检测点：量词“头” ■

■ **Example 1.5** His house is on the south bank of the river.

他的房子在河的南岸。

We keep our money in a bank.

我们在一家银行存钱。

检测点：bank 的多义翻译 ■

基于检测点的评价方法的意义在于，它并不是简单给出一个分数，而是帮助系统研发人员定位问题。因此这类方法更多的使用在对机器翻译的结果进行分析上，是对 BLEU 等整体评价指标的一种很好的补充。

1.6 机器翻译应用

机器翻译有着十分广泛的应用，下面看一下机器翻译生活中的具体应用形式：

（一）网页翻译

进入信息爆炸的时代之后，互联网上海量的数据随处可得，然而由于国家和地区语言的不同，网络上的数据也呈现出多语言的特性。当人们在遇到包含不熟悉语言的网页时，无法及时有效地获取其中的信息。因此，对不同语言的网页进行翻译是必不可少的一步。由于网络上网页的数量数不胜数，依靠人工对网页进行翻译是不切实际的，相反，机器翻译十分适合这个任务。目前，市场上有很多浏览器提供网页翻译的服务，极大地简化了人们从网络上获取不同语言信息的难度。

（二）科技文献翻译

在专利等科技文献翻译中，往往需要将文献翻译为英语或者其他语言，比如摘要翻译。以往这种翻译工作通常由人工来完成。由于对翻译结果的质量要求较高，因此要求翻译人员具有相关背景知识，这导致译员资源稀缺。特别是，近几年国内专利申请数不断增加，这给人工翻译带来了很大的负担。相比于人工翻译，机器翻译可以在短时间内完成大量的专利翻译，同时结合术语词典和人工校对等方式，可以保证专利的翻译质量。同时，以专利为代表的科技文献往往具有很强的领域性，针对各类领域文本进行单独优化，机器翻译的品质可以大大提高。因此，机器翻译在专利翻译等行业有十分广泛的应用前景。

（三）视频字幕翻译

随着互联网的普及，人们可以通过互联网接触到大量境外影视作品。由于人们可能没有相应的外语能力，通常需要专业人员对字幕进行翻译（如图 1.13）。因此，这些境外视频的传播受限于字幕翻译的速度和准确度。现在的一些视频网站在使用语音识别为视频生成源语字幕的同时，通过机器翻译技术为各种语言的受众提供质量尚可的目标语言字幕，这种方式为人们提供了极大的便利。



图 1.13: 电影字幕

（四）社交

社交是人们的重要社会活动。人们可以通过各种各样的社交软件做到即时通讯，进行协作或者分享自己的观点。然而受限于语言问题，人们的社交范围往往不会超出自己所掌握的语种范围，很难方便地进行跨语言社交。随着机器翻译技术的发展，越来越多的社交软件开始支持自动翻译，用户可以轻易地将各种语言的内容翻译成自己的母语，方便了人们的交流，让语言问题不再成为社交的障碍。

（五）同声传译

在一些国际会议中，与会者来自许多不同的国家，为了保证会议的流畅，通常需要专业译员进行同声传译。同声传译需要在不打断演讲的同时，不间断的将讲话内容进行口译，对翻译人员的素质要求极高，成本高昂。现在，一些会议开始采用语音识别来将语音转换成文本，同时使用机器翻译技术进行翻译的方式，达到同步翻译的目的。这项技术已经得到了多个企业的关注，并在很多重要会议上进行尝试，取得了很好的反响。不过同声传译达到真正的使用还需一定时间的打磨，特别是会议场景下，准确进行语音识别和翻译仍然具有一定挑战。

（六）医药领域翻译

在医药领域中，从药品研发、临床试验到药品注册，都有着大量的翻译需求。比如，在新药注册阶段，限定申报时间的同时，更是对翻译质量有着极高的要求。由于医药领域专业词汇量庞大、单词冗长复杂、术语准确且文体专业性强，翻译难度明显高于其他领域，人工翻译的方式代价大且很难满足效率的要求。为此，机器翻译近几年在医药领域取得广泛应用。在针对医药领域进行优化后，机器翻译质量可以很好地满足翻译的要求。

（七）中国传统语言文化的翻译

中国几千年的历史留下了极为宝贵的文化遗产，而其中，文言文作为古代书面语，具有言文分离、行文简练的特点，易于流传。由于其言文分离的特点，使得文言文和现在的标准汉语具有一定的区别，为了更好发扬中国传统文化，需要对文言文进行翻译。而文言文古奥难懂，需要人们具备一定的文言文知识背景，否则很难准确翻译，机器翻译技术也可以帮助人们快速完成文言文的翻译。除此之外，机器翻译技术同样可以用于古诗生成和对联生成等任务。

（八）全球化

在经济全球化的今天，很多企业都有国际化的需求，企业员工或多或少地会遇到一些跨语言阅读和交流的情况，比如阅读进口产品的说明书，跨国公司之间的邮件、说明文件等等。相比于成本较高的人工翻译，机器翻译往往是一种很好的选择。在一些质量要求不高的翻译场景中，机器翻译可以得到应用。

（九）翻译机

出于商务、学术交流或者旅游的目的，人们在出国时会面临着跨语言交流的问题。近几年，随着出境人数的增加，不少企业推出了翻译机产品。通过结合机器翻译、语音识别和图像识别技术，翻译机实现了图像翻译和语音翻译的功能。用户可以很便捷地获取一些外语图像文字和语音信息，同时可以通过翻译机进行对话，降低跨语言交流门槛。

（十）翻译结果后编辑

翻译结果后编辑是指在机器翻译的结果之上，通过少量的人工编辑来进一步完善机器译文。在传统的人工翻译过程中，翻译人员完全依靠人工的方式进行翻译，这虽然保证了翻译质量，但是时间成本高。相对应的，机器翻译具有速度快和成本低的优势。在一些领域，目前的机器翻译质量已经可以很大程度上减小翻译人员的工作量，翻译人员可以在机器翻译的辅助下，花费相对较小的代价来完成翻译。

1.7 开源项目与评测

从实践的角度，机器翻译的发展主要可以归功于两方面的推动作用：开源系统和评测。开源系统通过代码共享的方式使得最新的研究成果可以快速传播，同时实验结果可以复现。而评测比赛，使得各个研究组织的成果可以进行科学的对比，共同推动机器翻译的发展与进步。此外，开源项目也促进了不同团队之间的协作，让研究人员在同一个平台上集中力量攻关。

1.7.1 开源机器翻译系统

下面列举一些优秀的开源机器翻译系统：

统计机器翻译开源系统

- **NiuTrans.SMT**: NiuTrans[33] 是由东北大学自然语言处理实验室自主研发的统计机器翻译系统, 该系统可支持基于短语的模型、基于层次短语的模型以及基于句法的模型。由于使用 C++ 语言开发, 所以该系统运行时间快, 所占存储空间少。系统中内嵌有 n -gram 语言模型, 故无需使用其他的系统即可对完成语言建模。网址: <http://opensource.niutrans.com/smt/index.html>
- **Moses**: Moses[15] 是统计机器翻译时代最著名的系统之一, (主要) 由爱丁堡大学的机器翻译团队开发。最新的 Moses 系统支持很多的功能, 例如, 它既支持基于短语的模型, 也支持基于句法的模型。Moses 提供因子化翻译模型 (Factored Translation Model), 因此该模型可以很容易地对不同层次的信息进行建模。此外, 它允许将混淆网络和字格作为输入, 可缓解系统的 1-best 输出中的错误。Moses 还提供了很多有用的脚本和工具, 被机器翻译研究者广泛使用。网址: <http://www.statmt.org/moses/>
- **Joshua**: Joshua[18] 是由约翰霍普金斯大学的语言和语音处理中心开发的层次短语翻译系统。由于 Joshua 是由 Java 语言开发, 所以它在不同的平台上运行或开发时具有良好的可扩展性和可移植性。Joshua 也是使用非常广泛的开源机器翻译系统之一。网址: <http://joshua.sourceforge.net/Joshua/Welcome.html>
- **SilkRoad**: SilkRoad 是由五个国内机构 (中科院计算所、中科院软件所、中科院自动化所、厦门大学和哈尔滨工业大学) 联合开发的基于短语的统计机器翻译系统。该系统是中国乃至亚洲地区第一个开源的统计机器翻译系统。SilkRoad 支持多种解码器和规则提取模块, 这样可以组合成不同的系统, 提供多样的选择。网址: <http://www.nlp.org.cn/project/project.php?projid=14>
- **SAMT**: SAMT[35] 是由卡内基梅隆大学机器翻译团队开发的语法增强的统计机器翻译系统。SAMT 在解码的时候使用目标树来生成翻译规则, 而不严格遵守目标语言的语法。SAMT 的一个亮点是它提供了简单但高效的方式在机器翻译中使用句法信息。由于 SAMT 在 hadoop 中实现, 它可受益于大数据集的分布式处理。网址: <http://www.cs.cmu.edu/zollmann/samt/>
- **HiFST**: HiFST[10] 是剑桥大学开发的统计机器翻译系统。该系统完全基于有限状态自动机实现, 因此非常适合对搜索空间进行有效的表示。网址: <http://ucam-smt.github.io/>
- **cdec**: cdec[6] 是一个强大的解码器, 是由 Chris Dyer 和他的合作者们一起开发。cdec 的主要功能是它使用了翻译模型的一个统一的内部表示, 并为结构预测问题的各种模型和算法提供了实现框架。所以, cdec 也可以在被用来做一个对齐系统或者一个更通用的学习框架。此外, 由于使用 C++ 语言编写, cdec 的运行速度较快。网址: <http://cdec-decoder.org/index.php?title=MainPage>

- **Phrasal**: Phrasal[4] 是由斯坦福大学自然语言处理小组开发的系统。除了传统的基于短语的模型，Phrasal 还支持基于非层次短语的模型，这种模型将基于短语的翻译延伸到非连续的短语翻译，增加了模型的泛化能力。网址：<http://nlp.stanford.edu/phrasal/>
- **Jane**: Jane[31] 是一个基于短语和基于层次短语的机器翻译系统，由亚琛工业大学的人类语言技术与模式识别小组开发。Jane 提供了系统融合模块，因此可以非常方便的对多个系统进行融合。网址：<https://www-i6.informatik.rwth-aachen.de/jane/>
- **GIZA++**: GIZA++[22] 是 Franz Och 研发的用于训练 IBM 模型 1-5 和 HMM 单词对齐模型的工具包。在早期，GIZA++ 是所有统计机器翻译系统中词对齐的标配工具。网址：<https://github.com/moses-smt/giza-pp>
- **FastAlign**: FastAlign[5] 是一个快速，无监督的词对齐工具，由卡内基梅隆大学开发。网址：https://github.com/clab/fast_align

神经机器翻译开源系统

- **GroundHog**: GroundHog[1] 基于 Theano[25] 框架，由蒙特利尔大学 LISA 实验室使用 Python 语言编写的一个框架，旨在提供灵活而高效的方式来实现复杂的循环神经网络模型。它提供了包括 LSTM 在内的多种模型。Bahdanau 等人 在此框架上又编写了 GroundHog 神经机器翻译系统。该系统也作为了很多论文的基线系统。网址：<https://github.com/lisa-groundhog/GroundHog>
- **Nematus**: Nematus[26] 是英国爱丁堡大学开发的，基于 Theano 框架的神经机器翻译系统。该系统使用 GRU 作为隐层单元，支持多层网络。Nematus 编码端有正向和反向的编码方式，可以同时提取源语句子中的上下文信息。该系统的一个优点是，它可以支持输入端有多个特征的输入（例如词的词性等）。网址：<https://github.com/EdinburghNLP/nematus>
- **ZophRNN**: ZophRNN[36] 是由南加州大学的 Barret Zoph 等人使用 C++ 语言开发的系统。Zoph 既可以训练序列表示模型（如语言模型），也可以训练序列到序列的模型（如神经机器翻译模型）。当训练神经机器翻译系统时，ZophRNN 也支持多源输入。网址：https://github.com/isi-nlp/Zoph_RNN
- **Fairseq**: Fairseq[23] 是由 Facebook 开发的，基于 PyTorch 框架的用以解决序列到序列问题的工具包，其中包括基于卷积神经网络、基于循环神经网络、基于 Transformer 的模型等。Fairseq 是当今使用最广泛的神经机器翻译开源系统之一。<https://github.com/facebookresearch/fairseq>
- **Tensor2Tensor**: Tensor2Tensor[30] 是由谷歌推出的，基于 TensorFlow 框架的开

源系统。该系统基于 Transformer 模型，因此可以支持大多数序列到序列任务。得益于 Transformer 的网络结构，系统的训练速度较快。现在，Tensor2Tensor 也是机器翻译领域广泛使用的开源系统之一。网址：<https://github.com/tensorflow/tensor2tensor>

- **OpenNMT**: OpenNMT[12] 系统是由哈佛大学自然语言处理研究组开源的，基于 Torch 框架的神经机器翻译系统。OpenNMT 系统的早期版本使用 Lua 语言编写，现在也扩展到了 TensorFlow 和 PyTorch，设计简单易用，易于扩展，同时保持效率和翻译精度。网址：<https://github.com/OpenNMT/OpenNMT>
- **斯坦福神经机器翻译开源代码库**: 斯坦福大学自然语言处理组（Stanford NLP）发布了一篇教程，介绍了该研究组在神经机器翻译上的研究信息，同时实现了多种翻译模型 [19]。网址：<https://nlp.stanford.edu/projects/nmt/>
- **THUMT**: 清华大学 NLP 团队实现的神经机器翻译系统，支持 Transformer 等模型 [34]。该系统主要基于 Tensorflow 和 Theano 实现，其中 Theano 版本包含了 RNNsearch 模型，训练方式包括 MLE（Maximum Likelihood Estimate），MRT（Minimum Risk Training），SST（Semi-Supervised Training）。Tensorflow 版本实现了 Seq2Seq, RNNsearch, Transformer 三种基本模型。网址：<https://github.com/THUNLP-MT/THUMT>
- **NiuTrans.NMT**: 由小牛翻译团队基于 NiuTensor 实现的神经机器翻译系统。支持循环神经网络、Transformer 等结构，并支持语言建模、序列标注、机器翻译等任务。支持机器翻译 GPU 与 CPU 训练及解码。其小巧易用，为开发人员提供快速二次开发基础。此外，NiuTrans.NMT 已经得到了大规模应用，形成了支持 187 种语言翻译的小牛翻译系统。网址：<http://opensource.niutrans.com/niutensor/index.html>
- **MARIAN**: 主要由微软翻译团队搭建 [11]，其使用 C++ 实现的用于 GPU/CPU 训练和解码的引擎，支持多 GPU 训练和批量解码，最小限度依赖第三方库，静态编译一次之后，复制其二进制文件就能在其他平台使用。网址：<https://marian-nmt.github.io/>
- **Sockeye**: 由 Awslabs 开发的神经机器翻译框架 [9]。其中支持 RNNSearch、Transformer、CNN 等翻译模型，同时提供了从图片翻译到文字的模块以及 WMT 德英新闻翻译、领域适应任务、多语言零资源翻译任务的教程。网址：<https://awslabs.github.io/sockeye/>
- **CytonMT**: 由 NICT 开发的一种用 C++ 实现的神经机器翻译开源工具包 [32]。主要支持 Transformer 模型，并支持一些常用的训练方法以及解码方法。网址：<https://github.com/arthurxlu/cytonMt>
- **OpenSeq2Seq**: 由 NVIDIA 团队开发的 [17] 基于 Tensorflow 的模块化架构，用

于序列到序列的模型，允许从可用组件中组装新模型，支持混合精度训练，利用 NVIDIA Volta Turing GPU 中的 Tensor 核心，基于 Horovod 的快速分布式训练，支持多 GPU，多节点多模式。网址：<https://nvidia.github.io/OpenSeq2Seq/html/index.html>

- NMTPyTorch: 由勒芒大学语言实验室发布的基于序列到序列框架的神经网络翻译系统 [3], Numpytorch 的核心部分依赖于 Numpy, Pytorch 和 tqdm。其允许训练各种端到端神经体系结构，包括但不限于神经机器翻译，图像字幕和自动语音识别系统。网址：<https://github.com/lium-lst/nmtpytorch>

1.7.2 常用数据集及公开评测任务

机器翻译相关评测主要有两种组织形式，一种是由政府及国家相关机构组织，权威性高。如由美国国家标准技术研究所组织的 NIST 评测、日本国家科学咨询系统中心主办的 NACSIS Test Collections for IR (NTCIR) PatentMT、日本科学振兴机构 (Japan Science and Technology Agency, 简称 JST) 等组织联合举办的 Workshop on Asian Translation (WAT) 以及国内由中文信息学会主办的全国机器翻译大会 (China Conference on Machine Translation, 简称 CCMT)；另一种是由相关学术机构组织，具有领域针对性的特点，如倾向新闻领域的 Workshop on Statistical Machine Translation (WMT) 以及面向口语的 International Workshop on Spoken Language Translation (IWSLT)。下面将针对上述评测进行简要介绍。

- CCMT (全国机器翻译大会)，前身为 CWMT (全国机器翻译研讨会) 是国内机器翻译领域的旗舰会议，自 2005 年起已经组织多次机器翻译评测，对国内机器翻译相关技术的发展产生了深远影响。该评测主要针对汉语、英语以及国内的少数民族语言 (蒙古语、藏语、维吾尔语等) 进行评测，领域包括新闻、口语、政府文件等，不同语言方向对应的领域也有所不同。评价方式不同届略有不同，主要采用自动评价的方式，自 CWMT 2013 起则针对某些领域增设人工评价。自动评价的指标一般包括 BLEU-SBP、BLEU-NIST、TER、METEOR、NIST、GTM、mWER、mPER 以及 ICT 等，其中以 BLEU-SBP 为主，汉语为目标语的翻译采用基于字符的评价方式，面向英语的翻译采用基于词的评价方式。每年该评测吸引国内外近数十家企业及科研机构参赛，业内认可度极高。关于 CCMT 的更多信息可参考官网：<http://www.ai-ia.ac.cn/cwmt2015/evaluation.html> (链接为 CWMT 2015)。
- WMT 由 Special Interest Group for Machine Translation (SIGMT) 主办，会议自 2006 年起每年召开一次，是一个涉及机器翻译多种任务的综合性会议，包括多领域翻译评测任务、质量评价任务以及其他与机器翻译的相关任务 (如文档对齐评测等)。现在 WMT 已经成为机器翻译领域的旗舰评测任务，很多研究工作都以 WMT 任务作为基准。WMT 评测涉及的语言范围较广，包括英语、德语、

芬兰语、捷克语、罗马尼亚语等十多种语言，翻译方向一般以英语为核心，探索英语与其他语言之间的翻译性能，领域包括新闻、信息技术、生物医学。最近，也增加了无指导机器翻译等热门问题。WMT 在评价方面类似于 CCMT，也采用人工评价与自动评价相结合的方式，自动评价的指标一般为 BLEU、TER 等。此外，WMT 公开了所有评测数据，因此也经常被机器翻译相关人员所使用。更多 WMT 的机器翻译评测相关信息可参考官网：<http://www.sigmt.org/>。

- NIST 机器翻译评测始于 2001 年，是早期机器翻译公开评测中颇具代表性的任务，现在 WMT 和 CCMT 很多任务的设置也大量参考了当年 NIST 评测的内容。NIST 评测由美国国家标准技术研究所主办，作为美国国防高级计划署（DARPA）中 TIDES 计划的重要组成部分。早期，NIST 评测主要评价阿拉伯语和汉语等语言到英语的翻译效果，评价方法一般采用人工评价与自动评价相结合的方式。人工评价采用 5 分制评价。自动评价使用多种方式，包括 BLEU, METEOR, TER 以及 HyTER。此外 NIST 从 2016 年起开始对稀缺语言资源技术进行评估，其中机器翻译作为其重要组成部分共同参与评测，评测指标主要为 BLEU。除对机器翻译系统进行评测之外，NIST 在 2008 和 2010 年对于机器翻译的自动评价方法（MetricsMaTr）也进行了评估，以鼓励更多研究人员对现有评价方法进行改进或提出更加贴合人工评价的方法。同时 NIST 评测所提供的数据集由于数据质量较高受到众多科研人员喜爱，如 MT04, MT06 等（汉英）平行语料经常被科研人员在实验中使用。不过，近几年 NIST 评测已经停止。更多 NIST 的机器翻译评测相关信息可参考官网：<https://www.nist.gov/programs-projects/machine-translation>。
- 从 2004 年开始举办的 IWSLT 也是颇具特色的机器翻译评测，它主要关注口语相关的机器翻译任务，测试数据包括 TED talks 的多语言字幕以及 QED 教育讲座影片字幕等，语言涉及英语、法语、德语、捷克语、汉语、阿拉伯语等众多语言。此外在 IWSLT 2016 中还加入了对于日常对话的翻译评测，尝试将微软 Skype 中一种语言的对话翻译成其他语言。评价方式采用自动评价的模式，评价标准和 WMT 类似，一般为 BLEU 等指标。另外，IWSLT 除了对文本到文本的翻译评测外，还有自动语音识别以及语音转另一种语言的文本的评测。更多 IWSLT 的机器翻译评测相关信息可参考官网：<https://workshop2016.iwslt.org/>（链接为 IWSLT2016）
- 日本举办的机器翻译评测 WAT 是亚洲范围内的重要评测之一，由日本科学振兴机构（JST）、情报通信研究机构（NICT）等多家机构共同组织，旨在为亚洲各国之间交流融合提供便宜之处。语言方向主要包括亚洲主流语言（汉语、韩语、印地语等）以及英语对日语的翻译，领域丰富多样，包括学术论文、专利、新闻、食谱等。评价方式包括自动评价（BLEU、RIBES 以及 AMFM 等）以及人工评价，其特点在于对于测试语料以段落为单位进行评价，考察

其上下文关联的翻译效果。更多 WAT 的机器翻译评测相关信息可参考官网：<http://lotus.kuee.kyoto-u.ac.jp/WAT/>。

- NTCIR 计划是由日本国家科学咨询系统中心策划主办的，旨在建立一个用在自然语言处理以及信息检索相关任务上的日文标准测试集。在 NTCIR-9 的和 NTCIR-10 中开设的 Patent Machine Translation (PatentMT) 任务主要针对专利领域进行翻译测试，其目的在于促进机器翻译在专利领域的发展和应用。在 NTCIR-9 中，评测方式采取人工评价与自动评价相结合，以人工评价为主导。人工评价主要根据准确度和流畅度进行评估，自动评价采用 BLEU、NIST 等方式进行。NTCIR-10 评价方式在此基础上增加了专利审查评估、时间评估以及多语种评估，分别考察机器翻译系统在专利领域翻译的实用性、耗时情况以及不同语种的翻译效果等。更多 NTCIR 评测相关信息可参考官网：<http://research.nii.ac.jp/ntcir/index-en.html>

以上评测数据大多可以从评测网站上下载，此外部分数据也可以从 LDC (Linguistic Data Consortium) 上申请，网址为<https://www.ldc.upenn.edu/>。ELRA (European Language Resources Association) 上也有一些免费的语料库供研究使用，其官网为<http://www.elra.info/>。更多机器翻译的语料信息可参看附录??。

从机器翻译发展的角度看，这些评测任务给相关研究提供了基准数据集，使得不同的系统都可以在同一个环境下进行比较和分析，进而建立了机器翻译研究所需要的实验基础。此外，公开评测也使得研究者可以第一时间了解机器翻译研究的最新成果，比如，有多篇 ACL 会议最佳论文的灵感就来自当年参加机器翻译评测任务的系统。

1.8 推荐学习资源

首先，推荐一本书《Statistical Machine Translation》[13]，其作者是机器翻译领域著名学者 Philipp Koehn 教授。该书是机器翻译领域内的经典之作，介绍了统计机器翻译技术的进展。该书从语言学和概率学两个方面介绍了统计机器翻译的构成要素，然后介绍了统计机器翻译的主要模型：基于词、基于短语和基于树的模型，以及机器翻译评价、语言建模、判别式训练等方法。此外，作者在该书的最新版本中增加了神经机器翻译的章节，方便研究人员全面了解机器翻译的最新发展趋势 [14]。

《Foundations of Statistical Natural Language Processing》[20] 中文译名《统计自然语言处理基础》，作者是自然语言处理领域的权威 Chris Manning 教授和 Hinrich Schütze 教授。该书对统计自然语言处理方法进行了全面介绍。书中讲解了统计自然语言处理所需的语言学和概率论基础知识，介绍了机器翻译评价、语言建模、判别式训练以及整合语言学信息等基础方法。其中也包含了构建自然语言处理工具所需的基本理论和算法，提供了对数学和语言学基础内容广泛而严格的覆盖，以及统计方法的详细讨论。

《统计自然语言处理》[38]由中国科学院自动化所宗成庆教授所著。该书中系统介绍了统计自然语言处理的基本概念、理论方法和最新研究进展，既有对基础知识和理论模型的介绍，也有对相关问题的研究背景、实现方法和技术现状的详细阐述。可供从事自然语言处理、机器翻译等研究的相关人员参考。

Ian Goodfellow、Yoshua Bengio, Aaron Courville 三位机器学习领域的学者所写的《Deep Learning》[8]也是值得一读的参考书。其讲解了有关深度学习常用的方法，其中很多都会在深度学习模型设计和使用中用到。同时在该书的应用一章中也简单讲解了神经机器翻译的任务定义和发展过程。

《Neural Network Methods for Natural Language Processing》[7]是Yoav Goldberg编写的面向自然语言处理的深度学习参考书。相比《Deep Learning》，该书聚焦在自然语言处理中的深度学习方法，内容更加易读，非常适合刚入门自然语言处理及深度学习应用的人员参考。

《机器学习》[37]由南京大学教授周志华教授所著，作为机器学习领域入门教材，该书尽可能地涵盖了机器学习基础知识的各个方面，试图尽可能少地使用数学知识介绍机器学习方法与思想。

《统计学习方法》[39]由李航博士所著，该书对机器学习的有监督和无监督等方法进行了全面而系统的介绍。可以作为梳理机器学习的知识体系，同时了解相关基础概念的参考读物。

《神经网络与深度学习》[41]由复旦大学邱锡鹏教授所著，全面的介绍了神经网络和深度学习的基本概念和常用技术，同时涉及了许多深度学习的前沿方法。该书适合初学者阅读，同时又不失为一本面向专业人士的参考书。

TensorFlow 官网提供了一个有关神经机器翻译的教程，介绍了从数据处理开始如何利用 TensorFlow 工具从零搭建一个神经机器翻译系统以及如何解码，其地址为https://www.tensorflow.org/tutorials/text/nmt_with_attention。此外谷歌和 Facebook 也分别提供了基于序列到序列机器翻译模型的高级教程。谷歌的版本是基于 TensorFlow 实现，网址为：<https://github.com/tensorflow/nmt>，Facebook 的教程主要是基于 PyTorch 实现，网址为：https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html。网站上也包含一些综述论文，其中详细的介绍了神经机器翻译的发展历程，问题定义以及目前遇到的问题。

<http://www.statmt.org>是一个介绍机器翻译研究的网站，该网站包含了对统计机器翻译研究的一些介绍资料，一些自然语言处理的会议，常用工具以及语料库。<http://www.mt-archive.info>与<https://www.aclweb.org/anthology>网站上有许多介绍机器翻译和自然语言处理的论文。通过这个网站可以了解到自然语言处理领域的一些重要的期刊和会议。比如，一些与机器翻译相关的主要会议有：

- ACL, 全称 Annual Meeting of the Association for Computational Linguistics, 是自然语言处理领域最高级别的会议。每年举办一次，主题涵盖计算语言学的所有方向。

- NAACL, 全称 The North American Chapter of the Association for Computational Linguistics, 为 ACL 北美分会, 在自然语言处理领域也属于顶级会议, 每年会选择一个北美城市召开会议。
- EMNLP, 全称 Conference on Empirical Methods in Natural Language Processing, 自然语言处理另一个顶级会议之一, 由 ACL 当中对语言数据和经验方法有特殊兴趣的团体主办, 始于 1996 年。会议比较偏重于方法和经验性结果。
- COLING, 全称 International Committee on Computational Linguistics, 自然语言处理老牌顶级会议之一。该会议始于 1965 年, 是由 ICCL 国际计算语言学委员会主办。会议简称为 COLING, 是谐音瑞典著名作家 Albert Engström 小说中的虚构人物 Kolingen。COLING 每两年举办一次。
- EACL, 全称 European Chapter of the Association for Computational Linguistics, 为 ACL 欧洲分会, 虽然在欧洲召开, 会议也吸引了全世界的大量学者投稿并参会。
- AACL, 全称 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, 为 ACL 亚太地区分会。2020 年会议首次召开, 是亚洲地区自然语言处理领域最具影响力的会议之一。
- WMT, 全称 Workshop on Machine Translation。机器翻译领域一年一度研讨会, 是国际公认的顶级机器翻译赛事之一。
- AMTA, 全称 Association for Machine Translation in the Americas。AMTA 会议汇聚了学术界、产业界和政府的研究人员、开发人员和用户, 让工业界和学术界进行交流。
- CCL, 全称 Conference on Computational Linguistics, 中文为中国计算语言学大会。中国计算语言学大会创办于 1991 年, 由中国中文信息学会计算语言学专业委员会负责组织。经过 20 余年的发展, 中国计算语言学大会已成为国内自然语言处理领域权威性最高、规模和影响最大的学术会议。作为中国中文信息学会(国内一级学会)的旗舰会议, CCL 聚焦于中国境内各类语言的智能计算和信息处理, 为研讨和传播计算语言学最新学术和技术成果提供了最广泛的高层次交流平台。
- NLPCC, 全称 The CCF International Conference on Natural Language Processing and Chinese Computing。NLPCC 是由中国计算机学会(CCF)主办的 CCF 中文信息技术专业委员会年度学术会议, 专注于自然语言处理及中文处理领域的研究和应用创新。会议自 2012 年开始举办, 主要活动有主题演讲、论文报告、技术测评等多种形式。
- CCMT, 全称 The China Conference on Machine Translation, 中国机器翻译研讨

会，由中国中文信息学会主办，旨在为国内外机器翻译界同行提供一个平台，促进中国机器翻译事业。CCMT 不仅是国内机器翻译领域最具影响力、最权威的学术和评测活动，而且也代表着汉语与民族语言翻译技术的最高水准，对民族语言技术发展具有重要意义。

除了会议之外，《中文信息学报》、《Computational Linguistics》、《Machine Translation》、《Transactions of the Association for Computational Linguistics》、《IEEE/ACM Transactions on Audio, Speech, and Language Processing》、《ACM Transactions on Asian and Low Resource Language Information Processing》、《Natural Language Engineering》等期刊也发表了许多与机器翻译相关的重要论文。

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 2015.
- [2] Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. “A Statistical Approach to Machine Translation”. In: *Computational Linguistics* 16.2 (1990), pages 79–85.
- [3] Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. “NMTPY: A Flexible Toolkit for Advanced Neural Machine Translation Systems”. In: *Prague Bull. Math. Linguistics* 109 (2017), pages 15–28.
- [4] Daniel M. Cer, Michel Galley, Daniel Jurafsky, and Christopher D. Manning. “Phrasal: A Statistical Machine Translation Toolkit for Exploring New Model Features”. In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2, 2010, Los Angeles, California, USA - Demonstration Session*. The Association for Computational Linguistics, 2010, pages 9–12.
- [5] Chris Dyer, Victor Chahuneau, and Noah A. Smith. “A Simple, Fast, and Effective Reparameterization of IBM Model 2”. In: *Human Language Technologies: Confer-*

- ence of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA.* The Association for Computational Linguistics, 2013, pages 644–648.
- [6] Chris Dyer, Adam Lopez, Juri Ganitkevitch, Jonathan Weese, Ferhan Türe, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. “cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models”. In: *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*. The Association for Computer Linguistics, 2010, pages 7–12.
 - [7] Yoav Goldberg. “Neural network methods for natural language processing”. In: *Synthesis Lectures on Human Language Technologies* 10.1 (2017), pages 1–309.
 - [8] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016.
 - [9] Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. “Sockeye: A Toolkit for Neural Machine Translation”. In: *CoRR* abs/1712.05690 (2017).
 - [10] Gonzalo Iglesias, Adrià de Gispert, Eduardo Rodríguez Banga, and William J. Byrne. “Hierarchical Phrase-Based Translation with Weighted Finite State Transducers”. In: *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*. The Association for Computational Linguistics, 2009, pages 433–441.
 - [11] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. “Marian: Fast Neural Machine Translation in C++”. In: *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*. Association for Computational Linguistics, 2018, pages 116–121.
 - [12] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. “OpenNMT: Open-Source Toolkit for Neural Machine Translation”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations*. Association for Computational Linguistics, 2017, pages 67–72.
 - [13] Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.
 - [14] Philipp Koehn. “Neural Machine Translation”. In: *CoRR* abs/1709.07809 (2017).

- [15] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. “Moses: Open Source Toolkit for Statistical Machine Translation”. In: *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics, 2007.
- [16] Philipp Koehn, Franz Josef Och, and Daniel Marcu. “Statistical Phrase-Based Translation”. In: *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics, 2003.
- [17] Oleksii Kuchaiev, Boris Ginsburg, Igor Gitman, Vitaly Lavrukhin, Carl Case, and Paulius Micikevicius. “OpenSeq2Seq: extensible toolkit for distributed and mixed precision training of sequence-to-sequence models”. In: *CoRR* abs/1805.10387 (2018).
- [18] Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, Jonathan Weese, and Omar Zaidan. “Joshua: An Open Source Toolkit for Parsing-Based Machine Translation”. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation, WMT@EACL 2009, Athens, Greece, March 30-31, 2009*. Association for Computational Linguistics, 2009, pages 135–139.
- [19] Minh-Thang Luong and Christopher D. Manning. “Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.
- [20] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.
- [21] Makoto Nagao. “A framework of a mechanical translation between Japanese and English by analogy principle”. In: *Artificial and human intelligence* (1984), pages 351–354.
- [22] Franz Josef Och and Hermann Ney. “A Systematic Comparison of Various Statistical Alignment Models”. In: *Computational Linguistics* 29.1 (2003), pages 19–51.
- [23] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. “fairseq: A Fast, Extensible Toolkit for Sequence Modeling”. In: *Proceedings of the 2019 Conference of the North American Chapter of the*

- Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*. Association for Computational Linguistics, 2019, pages 48–53.
- [24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002, pages 311–318.
- [25] Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermüller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, Alexander Belopolsky, Yoshua Bengio, Arnaud Bergeron, James Bergstra, Valentin Bisson, Josh Bleecher Snyder, Nicolas Bouchard, Nicolas Boulanger-Lewandowski, Xavier Bouthillier, Alexandre de Brébisson, Olivier Breuleux, Pierre Luc Carrier, Kyunghyun Cho, Jan Chorowski, Paul F. Christiano, Tim Cooijmans, Marc-Alexandre Côté, Myriam Côté, Aaron C. Courville, Yann N. Dauphin, Olivier Delalleau, et al. “Theano: A Python framework for fast computation of mathematical expressions”. In: *CoRR* abs/1605.02688 (2016).
- [26] Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. “Nematus: a Toolkit for Neural Machine Translation”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Software Demonstrations*. Association for Computational Linguistics, 2017, pages 65–68.
- [27] Yu Shiwen. “Automatic evaluation of output quality for machine translation systems”. In: *Machine Translation 8.1-2* (1993), pages 117–126.
- [28] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. “A Study of Translation Edit Rate with Targeted Human Annotation”. In: *In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA. 2006*, pages 223–231.
- [29] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. “Sequence to Sequence Learning with Neural Networks”. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2014, pages 3104–3112.
- [30] Ashish Vaswani, Samy Bengio, Eugene Brevdo, François Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. “Tensor2Tensor for Neural Machine

- Translation”. In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*. Association for Machine Translation in the Americas, 2018, pages 193–199.
- [31] David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. “Jane: an advanced freely available hierarchical machine translation toolkit”. In: *Machine Translation* 26.3 (2012), pages 197–216.
- [32] Xiaolin Wang, Masao Utiyama, and Eiichiro Sumita. “CytonMT: an Efficient Neural Machine Translation Open-source Toolkit Implemented in C++”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*. Association for Computational Linguistics, 2018, pages 133–138.
- [33] Tong Xiao, Jingbo Zhu, Hao Zhang, and Qiang Li. “NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation”. In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*. The Association for Computer Linguistics, 2012, pages 19–24.
- [34] Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huan-Bo Luan, and Yang Liu. “THUMT: An Open Source Toolkit for Neural Machine Translation”. In: *CoRR* abs/1706.06415 (2017).
- [35] Andreas Zollmann, Ashish Venugopal, Matthias Paulik, and Stephan Vogel. “The Syntax Augmented MT (SAMT) System at the Shared Task for the 2007 ACL Workshop on Statistical Machine Translation”. In: *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*. Association for Computational Linguistics, 2007, pages 216–219.
- [36] Barret Zoph, Ashish Vaswani, Jonathan May, and Kevin Knight. “Simple, Fast Noise-Contrastive Estimation for Large RNN Vocabularies”. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. The Association for Computational Linguistics, 2016, pages 1217–1222.
- [37] 周志华. 机器学习. 清华大学出版社, 2016.
- [38] 宗成庆. 统计自然语言处理. 清华大学出版社, 2013.
- [39] 李航. 统计学习方法. 清华大学出版社, 2012.
- [40] 姚恺璇 赵军峰. “深化改革探讨创新推进发展——全国翻译专业学位研究生教育 2019 年会综述”. In: 中国翻译 (2019).

- [41] 邱锡鹏. 神经网络与深度学习. <https://nndl.github.io/>. 2020.

n -gram Precision, 28

n -gram 准确率, 28

n 元语法单元, 27

Brevity Penalty, 28

Data-Driven, 17

encoder-decoder, 24

Machine Translation, 12

Quality Estimation, 26

Quality Evaluation, 26

Source Language, 11

Target Language, 11

数据驱动, 17

无参考答案的评价, 26

机器翻译, 12

源语言, 11

目标语言, 11

短句惩罚因子, 28

编码器-解码器, 24

质量评价, 26