# 第八章

# 定制化机器翻译

Gema Ramírez-Sánchez
Prompsit Language Engineering

本章概述了以满足特定要求为目的的定制化机器翻译的理论意义和实践意义。本章的目标读者为机器翻译的初识者，不过对于想要向外行人士解释机器翻译的专业也会有所帮助。本章假设的机器翻译范式是神经网络机器翻译范式。

## 1 引言

### 1.1 通用机器翻译

毫无疑问，大多数偶尔使用机器翻译的用户都依赖于“通用”机器翻译，即机器翻译引擎的训练涵盖广泛的主题、文体和体裁，而非专门针对特定领域。

通用型引擎虽然很适合一般用途，但它很难用于词汇范围小、风格特殊，或受特定体裁惯例限制的文本。这种情况多见于法律或医学等专业化程度高的领域的文本，不过在日常生活中，我们也会见到这种文本特点。例如，食谱的特殊行文结构和用词有别于使用说明书等其他“日常”文本。例如，与使用说明书不同，食谱很少使用疑问句。这两类文本都经常被译成其他语言，或供日常使用（如搜索引擎翻译“如何做巧克力饼干?”、“要节能的话，推荐使用哪种灯泡?”），或有专业用途（如出版商要翻译食谱，或者制造商翻译消费品的技术规格说明）。在下文中，我会以食谱（苹果酥）和消费者指南（灯泡）为例，来展示通用机器翻译引擎如何处理这两类常见文本的术语。

食谱通常包括三部分：标题、食材和做法。对于通用型机翻引擎来说，光是简简单单的标题就足以叫人犯难。表 1 展示了三个通用型机翻系统给出的“apple crumble”（苹果酥）在西班牙语、法语和意大利语三种语言中的译法。从表中可以看出，机翻 1 给出的法语和意大利语译文尚佳，但西班牙语译文较差；机翻 2 的三种译文都不尽如人意；机翻 3 的法语译文不错，但西班牙语和意大利语的译文都不理想。

Gema Ramírez-Sánchez. 2025. 定制化机器翻译. In Dorothy Kenny (ed.), 机器翻译知识普及：为人工智能时代的用户赋能, 129–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922299

表 1: 食谱标题的机器翻译：apple crumble

| | 机翻 1 | 机翻 2 | 机翻 3 |
|---|---|---|---|
| 西班牙语 | migas<br>de manzana<br>‘crumbs of<br>apple’ | se desmorona<br>la manzana<br>‘the apple<br>falls apart’ | Desmoronamiento<br>de la manzana<br>‘falling apart of<br>the apple’ |
| 法语 | crumble<br>aux pommes<br>‘apple crumble’ | Crumble<br>d’ apple<br>‘crumble of apple’ | Crumble<br>aux pommes<br>‘apple crumble’ |
| 意大利语 | Crumble<br>di mele<br>‘apple<br>crumble’ | La mela<br>si sbriciola<br>‘the apple<br>crumbles’ | Crumble<br>di mela<br>‘crumble<br>of apple’ |

不懂西班牙语、法语或意大利语的话，也没关系。你可以选一个喜欢的在线机翻引擎，把“apple crumble”译成自己懂的语言。得到的译文可能类似表表 1中的 collapsing apples 或 apples that fall apart。其他译文都很好，如 crumble aux pommes 和 Crumble di mele（这就是为什么我们在表 1中只简单将其注释为 apple crumble）。机翻引擎输出的译文取决于其训练数据。在此基础上，我们还可以添加一些步骤，如在训练过程或译后编辑阶段给引擎提供正确的术语。但是，如果不给引擎提供术语，通用机翻引擎往往就会产出如上文中机翻 2 的低质输出。

下面，我们一起来具体分析灯泡的例子。对于要想了解市面上各式各样灯泡的消费者来说，有大量相关的专业信息可供参考。假设你是生活在英语国家的非英语母语者，有盏灯坏了，邻居想给你送个 twisted fluorescent lamp。你听不懂邻居说的话，只好用笑容回应。于是，你拿出手机，使用一些通用型机翻系统来查看翻译，得到的结果如表 2所示。

看完这些译文，你以为邻居会给自己一盏造型奇特的常见灯泡或台灯。这一猜测由何而来？原来，是因为译文表达模糊不清：所有机翻引擎都没有使用表示“灯泡”这一含义的词来翻译“lamp“，而都采用该词的另一个含义 lighting equipment（照明设备）。

通用机器翻译翻错了，不过定制化机器翻译引擎应该能翻对。那么，什么是定制化机器翻译呢？请看下节内容。

表 2: 某种类型的灯泡：螺旋形荧光灯泡。

| | 机翻 1 | 机翻 2 | 机翻 3 | 正确译文 |
|---|---|---|---|---|
| 西班牙语 | Lámpara fluorescente retorcida | Lámpara fluorescente retorcida | Lámpara fluorescente retorcida | Bombilla fluorescente en espiral |
| 法语 | Lampe fluorescente torsadée | Lampe fluorescente tordue | Lampe fluorescente torsadée | Ampoule spirale fluorescente |
| 意大利语 | Lampada fluorescente contorta | Lampada fluorescente contorto | Lampada fluorescente attorcigliata | Lampadina fluorescente spirale |

## 2 定制化机器翻译

与通用机器翻译相反，“定制化机器翻译”是为满足特定目的而设计的机器翻译。

假设你在一家生产大牌汽车的公司工作。和汽车行业的其他制造商一样，你的公司也要用几十种语言来产出大量技术手册、用户指南和营销资料。任何有助于公司提升与内部员工的沟通效率、培训汽车销售人员或吸引客户等，都是关键事项，在像这样的多语言环境中，机器翻译能够大显身手。因此，你公司使用机器翻译来翻译几乎所有资料的初稿，然后由审稿人，也就是所谓的译后编辑者（参见O’Brien 2025 [本卷]）来改善译文。

你的公司最开始先使用通用机器翻译系统，然后通过译后编辑来提高机器译文质量。译后编辑者很快便发现，他们需要一遍遍重复修改相同的术语、体裁和风格错误，这一过程不但枯燥无味且效率低下。然后，考虑到这样的资料翻译已经进行了几十年，公司想知道，是否有办法能利用现有的翻译数据来改善翻译流程。

答案是肯定的。但是怎么做呢？首先，公司要用自己以往的翻译数据来训练机翻引擎，也就是说，用自己的训练数据来打造定制化机翻引擎。定制化机翻引擎翻译的初稿更接近公司过去的翻译，术语和风格方面的错误也少得多，译后编辑者也更高兴。

但是，事情真的就这么简单吗？是的，不过前提是你有足够多的数据（数百万个翻译过的句子），数据格式正确（对齐的平行语料；参见Kenny 2025 [本卷]）、保持内部一致（否则产出译文不一致），且为所需的语言对。你还需要机器翻译算法工程师或外部供应商来训练系统，将其融入公司的翻译工作

流程以及合适的硬件和软件中。这一切还只是开始。然后，如果想继续利用新的翻译资料，就需要为系统重新训练进行规划：若采用自适应式机器翻译，这可以即时完成；若翻译产出量巨大，那么每六小时就需要重新训练一次系统；如果只为了更新或保持系统一致，则每六个月或一年一次。

所以，这一过程或许不能用“简单”来形容，你也许会想是否有必要如此大费周章？下面我们来看看对定制化机器翻译有哪些合理期待。

## 2.1 我们对定制化机器翻译的期待是什么？

定制化机器翻译曾专属于机器翻译专家，但如今已普遍为各种用户所用。我们甚至能看到定制化机器翻译输出的未经修订的原始译文，只需点击“获取翻译”便可一键生成。酒店的在线预订网站、专业软件的在线技术支持、招聘职位列表、教育应用程序中的教师信息等都可以看到定制化机器翻译的踪影。

作为普通用户，我们的主要目的通常是理解从某个渠道（通常是网站）检索到的信息。在这种情况下，我们可能希望至少在术语和习语方面，定制化机器翻译比通用机器翻译更准确，也希望能保持文本风格一致。

例如，上网查找疗养院的相关信息时，我们希望导航菜单中的“home”译为“主页”，而网页内容中出现的“home”译为“老年人照护机构”。而在浏览棒球网站时，我们期待定制化机器翻译能够区分“home”的不同含义。在其他语境中，“home”这个单词根本不应该翻译，如作为品牌名称的一部分。我们希望定制化机器翻译能更好地处理特定语境下的翻译。

对于语言专业人士来说，其他看似微小的细节尤为重要。他们希望当定制化机器翻译引擎可以正确输出单词的大小写形式、处理项目符号列表等格式惯例，或者恰当表示数字时，他们会感激不尽。反之（如表 3），译后编辑者就需要仔细检查机翻引擎译文，纠正这些“小”错误，这既令人恼火又耗时。在专业环境中，我们也希望机翻译文能符合文本规范和问题要求。

如果训练力度足够且资源合适，那么定制化机器翻译的输出不会出现表 3所示的错误。所需的资源包括合适的人力资源，我将在下一节讨论这一点。

## 2.2 谁来定制机器翻译？

我们会从这一小节了解到，机器翻译的确是个跨学科领域，需要计算机工程师与语言学家的共同努力。

在基于规则的机器翻译为领先技术的阶段，计算机工程师和语言学家都在机翻系统的搭建中发挥着积极的作用：语言学家编写语法规则和词典，计算机工程师则负责编写程序来执行规则。

表 3: 机器翻译输出的细节很重要

| 输入 | 输出：通用机器翻译 | | |
|---|---|---|---|
| 英语 | 西班牙语 | 法语 | 意大利语 |
| For this match, the following players will be excluded | Para este partido, los siguientes jugadores serán excluidos: | Pour ce match, les joueurs suivants seront exclus: | Per questa partita, saranno esclusi i seguenti giocatori: |
| a) One | uno | un | un solo |
| b) Four | b) cuatro | b) Quatre | b) Quattro |
| c) 6 | c) seis | c) Six | c) Sei |

后来，到了统计机器翻译时代，工程师几乎负责整个机器翻译系统的搭建工作。语言学家偶尔参与输出评估，但很少参与错误分析，也参与制定改进机器翻译输出的具体措施。

然而，在统计机器翻译时代，这项技术仍然依赖于译者，因为他们提供的翻译为翻译模型提供了训练数据。[1] 但我可以肯定地说，在统计机器翻译的全盛时期，语言学家或多或少被排除在机器翻译系统的实际搭建工作之外。

如今，精通语言学的工程师逐渐不仅开始关注参数调整和硬件或自动评测指标，还开始关注机翻系统在高细粒度级别输出的译文质量。与此同时，精通技术的语言学家也已经开始参与评测和处理训练引擎所需的数据，使用训练工具包和评测系统，并制定策略改善系统，思考如何将其融入客户公司的翻译工作流程中。

鉴于机器翻译的本质，理想的情况是这两个领域的专业人员开展合作，有兴趣了解彼此的领域并为之做出贡献。这对于定制化机器翻译特别有效，因为有了语言学家评测训练数据的有用性、调控（或至少理解）训练过程遵循的策略，以及分析输出，对定制特定领域的机翻系统大有裨益。定制化机器翻译还受益于工程师对他们正在使用的文本和语言具体细节的理解，以便创造性地找到解决方案来推进和解决译文存在的主要问题，例如是否需要新模块、预处理或后处理步骤来处理多样的词语形态、产品名称或字母数字编码？

越来越多的学校针对语言学家和工程师增设机器翻译技术或语言学方面的培训课程，但这类培训主要还是出现在需求更加灵活变动的职业场景。

在本节结束前，还有一个群体值得提及，他们对定制化机器翻译的贡献与日俱增，但不一定会被称作“语言学家”，那就是“译者”。在这一章中，我主

[1] 下次听到有人说“机器翻译正在实现人类平等而‘无需人类干预’”时，请记住这一点。若真是如此，用来训练引擎的文本从何而来呢？

要将定制化机器翻译作为一种离线活动来介绍，只是偶尔需要人工干预。然而，值得注意的是，目前有些“自适应式”机器翻译的设置（参见O'Brien 2025 [本卷]）可以实时进行定制化（或至少比其他设置更快且更频繁）。在这种情况下，译者及其译文正成为定制化机器翻译的基石，而定制化实则为实时进行的用户模仿。当译者提供新的平行句对（源语-目标语句对），新的翻译会自动保存到已有系统中，成为译者可用的首选译文。

# 3 如何定制机器翻译引擎

## 3.1 一则寓言

假设你降落在新的星球上，没有生命迹象，但有座宏伟的图书馆，里面似乎藏有大量不同语言的文本：有些用这种语言（L1），有些用另一种语言（L2），还有一些是双语文本（L1-L2）。但是，由于没有关于语法、拼写和系统的语言学知识的书，所以无从知晓这些语言的使用方式，眼前只有大量句子或句对。

你还偶然发现了双语词汇表。此外，在查看这些文本时，你发现有些带印章，印章还各不相同，还有些不带任何印章。[2] 你感到庆幸，因为很可能只有通过这些文本，才可以了解这个新发现星球失落的生命遗迹。

但事情并非如此！等等！就在即将离开图书馆时，你发现星球上还有生命——两个星球居住者恶狠狠地盯着你，但你也注意到他们之间也相互怒视。过了一会儿，你发现他们敌意主要源自无法相互理解，一个说 L1 语，另一个说 L2 语，且都不知道图书馆的存在。你需要帮助他们！作为一个通晓多种语言的人，你也曾提供这样的帮助。现在，你别无选择，只能教会他们如何互译这两种语言。

你要怎么做呢？先单独学习这两种语言，再深入研究怎么进行互译？（这么做很耗时。）还是直接从学习双语文本和单词表开始？（这种方法似乎更好。）你可以从单词表开始，仔细观察其他单词、短语和更长的语块之间的关系。或者，可以把文本分为带印章和不带印章两类，例如将带有相似印章的文本放在一起。又或者，可以同时使用所有文本。如何从这些数据中学习两种语言的互译，你有很多方法可供选择。

在这一点上，学习如何翻译的机器翻译系统和你面对同样的处境：都有双语（和单语）文本，可能还有术语表，但仅此而已。这些是唯一的学习资源，不过学习过程多种多样。

回到新发现的星球，你首先发散思维，尝试从双语文本寻找学习思路，然后通过已编制好的单词表来验证自己的假设，接着在这些假设的基础上做出新的假设。你的观察对象很快就从单词发展到更长的语块。在这一阶段，你并不考虑文本是否带有印章，而是同时使用所有资源。

[2] 在这个比喻中，不同印章代表不同领域，而没有印章的文本可被视为非特定领域的文本。

同样地，要新建一个机器翻译系统，通常先不分领域地把所有双语数据都串联起来，使用默认的软件设置进行初始训练。

在第一次训练后，你开始把 L1 语言使用者表达的信息译成 L2 语言，再让 L2 语言讲者来验证译文，然后在另一个翻译方向 (L2->L1) 执行同样的操作。你可以通过这两种语言使用者的面部表情来丰富语言知识。有时他们会哈哈大笑，但大多时候会点头示意，有时甚至看起来好像理解了你的译文。你从他们的反馈中学习并继续改进。

在机器翻译系统开发过程中，评测通常不是基于人类（或外星人）的评估。相反，我们使用自动评测指标来对比机翻系统译文和职业译者译文，从而计算出质量得分（参见Rossi & Carré 2025 [本卷]）。在大多数情况下，机器译文与人工翻译越相似，质量就越高。如果机器译文的自动评测结果不错，且快速审阅后没发现严重的翻译错误，则该系统便可视为功能基准。否则，我们需要继续训练系统，比如增加一些预处理或后期处理。每轮训练结束后，都要用自动评测指标来检查，直至结果达到我们满意的水平，才会停止训练。

回到这个新发现的星球，你的学习取得了进展，还发现对于同一语言组合，有些单词虽然不止一种译法，但在带有相同印章的文本中的译法始终保持一致。因此，你决定按照印章将文本归类，并为每个类别分别制作译法列表。接着，你开始先后观察不带印章和带印章的文本，这两类文本略有不同，例如有些带章文本往往使用长句，而不带章文本则使用短句。

在这种情况下，根据拥有的数据量和系统的最终使用目的，我们可以只使用带相同印章的文本（印章表示领域）来训练机器翻译引擎。我们的域内系统可以用一般领域和特定领域的文本，也可以只用特定领域的文本。我们一定会充分利用最先进的机器翻译技术，尽量打造针对特定领域的系统。这正是定制化意义所在——玩转数据和技术。我们将在下文用简单易懂的方式来解释这些方法。有关神经网络机器翻译的“领域自适应”的更全面的研究，可参考Saunders (2021)。

## 3.2 基于数据的定制

机器翻译系统可使用特定文本，来为特定用途进行调整，例如我们可以为移动电话的描述开发优质的机器翻译系统，只要拥有足够多的描述移动电话的文本和对应的目标语译文。我们也可以使用相关领域的单语文本或双语词汇表。这就是我们所说的“域内数据”（in-domain data）。理想的情况是，拥有语句平行对照[3] 的双语域内数据。

[3]即源语-目标语的翻译句对，最好按照在原文出现的顺序排列，以便利用邻近句子的语境（参见 Kenny 2025 [本卷]）。

### 3.2.1 我们需要多少数据?

很难说要多少数据才算够。对于通用机器翻译系统来说，能够获取的数据越多越好，然后基于某些质量标准进行筛选，例如删除多次重复的句子（可以学习的内容太少）、内容杂乱的句子（如大多为数字的句子），以及太长的句子（学习难度太大）。而对于定制化机器翻译系统，答案可能相同，但还要考虑到域内数据必须占整个训练数据集的很大一部分，否则系统无法学习如何产出域内翻译。

这里特地只宽泛地说“很大一部分”，是因为我们知道很少有足够的域内数据来训练系统。毕竟，我们至少需要几百万个句对。这与通用引擎所需数据相比较少，但依然是相当庞大的数据量。因此，我们通常会结合使用域内数据和通用或域外数据。

定制化机器翻译系统的第一步如果是把可用的域内数据添加到域外数据，根据不同的语言组合，我们通常会遇到不同情况：数据要么太多，要么太少。而数据的规模很重要。

#### 3.2.1.1 数据“过多”

数据过多时，系统训练会耗费大量时间，需要运行更多服务器，这样的操作有些不切实际。至于多少数据算多，很难给出确切的数字，假设你发现减少数据后结果相同，数据就算过多了。减少数据的好处在于，所需的计算资源和训练时间都更少。（进一步了解开发人员如何使用 BLEU 等指标来判断是否让引擎停止学习，请参见Pérez-Ortiz et al. (2025 [本卷]), 尤其是7.2）。

开发定制化机器翻译系统，要优先使用所有能获取的域内数据。而对于域外数据，我们则选择可用的平行句对子集。通常以域内数据作为理想模板来筛选数据。数据的自动筛选方法有很多种，如：

- 根据与域内句子的语篇、语义或句法的相似度，对域外句子进行评分
- 根据在域内数据中发现的主题，对域外数据进行分组
- 根据优质或低质的域内句子，把域外句子重新分类为域内句子

#### 3.2.1.2 数据“过少”

相反，数据过少时，如果无法获取更多的数据，则会有损机器翻译系统的性能。系统可能无法翻译很多词语，输出质量很差的译文。在这种情况下，首先要基于已有域内数据进行数据扩展。我们可以免费获取或考虑购买现成的数据库或定制数据。自动数据扩展的方法包括：

- 通过爬取包含目标语的多语言网站来获取更多双语数据，要格外留意域内数据所含词汇
- 使用更多的单语数据，最好是目标语，并通过第三方机器翻译系统将其回译至源语
- 使用中转语来翻译单语数据，首先将单语数据译为中转语，再译入我们所需语言对的目标语或源语
- 利用现有数据自动合成新句子（将单词替换成同义词或频率相同的词、采用自动改写等），从而创造出新的源语和目标语句子

经验表明，不仅双语数据，而且单语和多语数据，以及通用、域内和多域数据，都在帮助机器翻译系统学习方面发挥了作用（参见 (Saunders 2021)）。更重要的是，自适应或增量机器翻译都开始考虑使用少量数据（参见O'Brien 2025 [本卷]）。机器翻译的开发日新月异，但有一点是肯定的：只要有数据，就能学习，机器翻译系统会充分利用这些数据。

### 3.2.2 数据质量

数据的质量也会在机器翻译系统的定制中发挥作用，直接影响到最终输出的质量。尤其随着神经网络机器翻译的兴起（参见Kenny 2025 [本卷] 和 Pérez-Ortiz et al. 2025 [本卷]），数据的质量成为热门话题，因为有研究表明神经网络机器翻译对训练数据的噪声非常敏感 (Khayrallah & Koehn 2018)。大多数致力于解决这个问题的研究包括结合使用模式和规则来过滤明显的噪声、句子质量评分，以及根据内容质量将数据分为优质和低质两类。此外，还有研究采用删除重复内容的方法 (Khayrallah & Koehn 2018)。

### 3.2.3 数据的结构

最后，数据的结构也是热议话题 (Mohiuddin et al. 2022)。有些研究使用长度相似的句子作为训练数据，以提高训练和翻译速度，有些则通过给模型输入由易到难的句子来提高质量，还有些使用成篇文档，而不是打乱的句子，以便利用更大的语境来改善机器翻译质量。

## 3.3 基于技术的定制

整合清理了所有数据后，接下来可以进入域内机器翻译系统的训练环节。该如何使用这些数据？哪类系统架构最适用于我们的目的？我们有不同选项可用于约束输出吗？这些就是基于技术的定制所要讨论的内容。这些技术可能

涉及神经网络架构的调整、训练期间的参数调整，以及训练或翻译期间（也称为推理）不同系统的组合。

我会在下文介绍一些开发特定领域机器翻译系统的最常见技术。更多讨论可参考Koehn (2020: Ch. 13)。

### 3.3.1 自学系统

对于自学系统，我们最初仅用通用型数据来训练，然后以此为起点，仅用域内数据进行第二次训练。这种方法训练出来的微调系统既有一般性语言知识，也有域内词汇和结构等专业知识。

### 3.3.2 受导系统

除了使用通用和域内数据来训练机器翻译系统之外，还可以使用基于目标语域内文本训练而成的语言模型，不过起码是高质量的单语文本。该语言模型有助于生成更符合域内语言特点的文本。此外，也可以使用域控制器或鉴别器，给训练数据进行单词、句子甚至词嵌入级别的标记。具体来说，就是先精准识别哪些通用数据与域内数据相似或不同，然后利用这些信息进行训练。

### 3.3.3 合作系统

有些系统由若干部分组成（如特定领域的子神经网络）或甚至可以包含多个完整系统。它们结合所有系统的知识可以生成比单个系统质量更好的文本。

## 4 实践中的定制

理论和实践往往不可分割。本节会介绍神经网络机器翻译引擎定制化的实践细节，内容主要针对初学者，不涉及复杂的技术知识。

### 4.1 可用工具

机器翻译定制化的专业工具已有好几种可供选择，但大多只允许用户改变数据，不可以调整技术。因此，虽然很多语言专业人士都有能力进行基于数据的定制化，但基于技术的定制化仍然只能是机器翻译研究者和开发者的专业领域。

大多数机器翻译供应商以定价方式提供对预训练通用或特定领域机翻引擎的远程访问，有些还提供定制服务。[4] 定制服务通常包括：

[4]在撰写本文时，有 40 多家供应商提供机翻服务，约 20 家提供一些定制服务。来源：https://inten.to/mt-landscape/，最后访问于 2022 年 6 月 26 日。

- 添加用户的语料库
- 添加用户的术语
- 利用用户的数据训练新模型
- 测试定制化模型

以上内容都可以半自动方式完成，即人工处理数据和相关流程，或以全自动方式完成，即定制时无需人工干预。

还有专为教授语言专业人士如何使用机器翻译而设计的机器翻译测试环境。这些测试环境作为教学工具，用于翻译技术课堂或职业场景，通常提供定制化选项，以观测系统在使用通用或域内数据训练时有何变化，如 MutNMT 就是一个很好的例子。[5]

## 4.2 策略制定的关键因素

以下是机器翻译系统定制策略的最重要因素：

- 语言对
- 领域
- 可用数据
- 系统用途
- 定制期限
- 硬件特性和可用性

下面，我们来一一了解。

你可以根据语言组合来决定使用哪种工具。例如，形态变化丰富的语言适合使用支持预处理和后期处理的工具。

有些领域的语篇特点突出，需要特定训练工具包的支持。例如，某个特定训练工具包可能非常适合处理很长或很短的句子、数字表达，以及通常需要在译文中保留的专有名词。如果该领域还有其他特点，也需要一并考虑。

定制系统时，可用数据的数量和质量不仅决定训练的进度，还决定是否需要增加数据或提高数据质量。此外，如果想添加平行句对以外的数据，必须确保有技术支持。

[5]参见 http://www.multitrainmt.eu/index.php/en/neural-mt-training/mutnmt，最后访问于 2022 年 6 月 26 日。

大多数神经网络机器翻译系统都致力于产出最好的译文，但是这通常还需要满足其他要求。选择服务供应商前，可能需要先考虑其语言模型是否支持远程访问，或是否真的需要在本地进行访问。该系统是否可以与其他用户同时在线使用，还是需要排队？该系统是否能翻译文本字符串？还是需要用不同文件格式来获取翻译？是支持 API、网络应用程序，还是通过第三方工具的连接器进行访问？诸如此类问题。

机器翻译系统的定制规划可能需要权衡质量和交付期限。训练出最佳系统所需要的时间，可能远超你的预期。

最后，无论是系统的训练，还是其使用，硬件都是关键因素。综合考虑其他因素，你可能需要全天候的服务，且同时使用多个 GPU 或 CPU。

## 4.3 获取正确数据

机器翻译系统的定制在很大程度上依赖于域内平行数据的可用性和质量，且这些数据要接近需要翻译的原文。在3.2.1中，我讨论了数据规模，以及如何根据场景来选择或扩展数据的方法。在此，我想讨论一个很基础的问题：在哪里可以找到适合系统的平行数据？

首先，可以获取现有的免费平行数据。有很多公开的免费数据库提供可用于机器翻译的双语语料库。这类数据库或语料库大多可在由赫尔辛基大学维护的 OPUS 网站获取，该网站提供 200 多种语言组合的语料库。[6]

其次，可以购直接买平行数据（出售供机器翻译使用的平行数据也是一门生意）。待售的数据规模大小不一，使用权、商业模式和定价也不尽相同。TAUS 是最大的机器翻译数据集供应商，提供覆盖 600 多种语言对的平行语料。[7]

此外，还可以自行构建平行数据。虽然难以搜集到大量数据，但还是可以通过以下方式构建平行语料库：

- 爬取多语言网站：可以通过包含双语内容的 URL 下载所需语言对的双语文本，并自行对齐（对齐文本的例子可参见Kenny 2025 [本卷]），或使用第三方服务进行对齐。
- 对齐自己翻译过的文本：如果有翻译过的文档及其源语文本，可以使用标准的开放获取或专有工具进行句级对齐。[8]

[6]参见 https://opus.nlpl.eu/index.php，最后访问于 2022 年 6 月 26 日。

[7]参见 https://www.taus.net/，最后访问于 2022 年 6 月 26 日。

[8]免费对齐工具有 LF Aligner (https://sourceforge.net/projects/aligner/，最后访问于 2022 年 6 月 26 日)。有些为人熟知的付费对齐工具还提供翻译记忆库工具（参见 Kenny 2025 [本卷]）。

## 4.4 正确处理数据

获取正确数据后，我们需要在训练系统前进行数据准备，确保：

- 每个文件只含单语（或语言对）语料，因为机器翻译训练常使用纯文本文件。理想的情况是每个文件都是单语的，尽管有些系统可以处理分列对照的电子表格形式（每列一种语言），甚至 TMX 文件（参见Kenny 2025 [本卷]）。
- 每行一句：文本文件为每行一句，也就是说一句话单独成段。通常不允许出现多句或不成句的情况[9]，不过标题也可视为句子，独占一行。
- 句子行行对齐：无论哪种格式，源语文件与目标语文件都必须行行对齐。
- 数据清理干净：无重复、拼写错误或含噪音的句子（如句子仅含数字、出现乱码或出现其他语言）。
- 数据经过匿名化处理（如有必要）：要删除训练语料库中的任何敏感数据，尤其是个人数据。
- 数据结构清晰：语料库需按训练的不同阶段分为三个数据集，通常命名为“训练集”（training set，简称 train）、“验证集”（validation set，也称为“开发集”(development set，简称 dev)）和“测试集”（test set，或简称为“test”）。

建议在整理数据时应该：

- 必须避免这三个数据集出现重叠。其实，如果可能的话，每个数据集应使用来源不同的句子，以保证其平衡和独立性。
- 训练集的规模从几千到几百万个句子不等，但验证集和测试集通常不会超过 5000 句。
- 训练数据可能包含通用型和定制化数据，但验证集和测试集应尽量使用域内数据，以便测试训练模型是否能够成功翻译域内文本。
- 若训练集同时含有通用数据和域内数据，应尽量提高域内数据的比例，否则模型将主要从通用数据中获取知识。

至此，我们讨论过的内容都并非为定制化引擎特有。这些准备工作均适用于通用型和定制化机器翻译的训练。下面将要介绍的预处理相关步骤也是如此，但可能因特定语言或语言组合而异：

[9]此处的训练单位为句子，而非文档。

- 文本分词：需要为训练提供分词准确的文本。分词单元（token）指文本分成的不同单位，包括单词、空格或标点符号。分词主要是识别词的边界，有时需要判断词的起止位置，在笔语中通常以空格为界。但是，泰语等语言的词语之间没有空格，这就增加了分词的难度。有时，这类语言的分词器只是将句子切分为看似无明显词界限的字符序列。这种方法虽不考虑“意义单位”，但优点是适用于任何语言。

- 文本大小写的处理：我们或许希望系统通过训练学习到，在训练数据中，句首首字母大写单词（如 The）与该词在句子其他位置的小写形式（如 the）是同一个词。因此，我们使用大小写转换器（truecaser）将（在英语等语言中的）专有名词之外的所有单词转为小写。[10] 大小写转换仅适用于区分大小写的语言，因此不适用于中文、阿拉伯语、希伯来语等语言。

- 子词拆分：根据文本分词情况，还可以进一步将单词拆分为子词、字符或其他语块。这种拆分可以基于频数或具有语言学意义的单元，如语素、词干和附加的形态学信息等。但是，并非所有定制化系统都拥有支持这种方法的配置环境，因为它要求神经网络能够处理特殊类型的输入和输出，而预处理和后处理通常只支持一种方法。

那么，如果你想使用的训练数据格式不对，该怎么办呢？别着急，大多情况下，分词、大小写转换和子词拆分（若适用）均为标准训练工具包预训练、预处理和后期处理步骤的默认环节。之所以在此提到这些环节，是为了让用户对训练工具包有所了解。还有很多独立工具可以实现这些步骤。这类工具可通过 Hugging Face [11] 或 Github [12]等平台获取，或直接在搜索引擎上查找“句子切分器”（sentence splitter）、“句子对齐工具”（sentence aligner）、“平行语料库过滤”（parallel corpus filtering）、“匿名化工具”（anonymizer）、“分词器”（tokenizer）、“大小写转换器”（truecaser）等。要进行精准搜索，还可以添加源语和目标语。通过以上方法，你会发现很多工具。

## 4.5 训练定制化模型

获取格式正确的适用数据后，在目前的很多训练环境中，训练机器翻译模型只需点击一个按键。有些环境允许用户调整许多参数，以充分利用数据训练环境和系统架构。有时，用户可以根据教育或研究目的来调整设置。

在训练前，用户最常调整的参数如下：

[10]注意，在德语中，所有名词的首字母都要大写，因此这些名词跟英语专有名词一样，不应该进行大小写转换。

[11]https://huggingface.co/

[12]https://github.com/，最后访问于 2022 年 6 月 26 日

- “词表大小”(Vocabulary size) 规定从训练语料库中计算得出的不同单词或子词（也称子词单元、类型或单词类型）的数量。
- “批量大小”(Batch size) 指每个训练步骤一并处理的分词单位[13]的数量。这一步很有必要，因为不可能将训练集的所有数据一次性输入神经网络。
- “波束大小”(Beam size) 是在翻译一个单词时，需要考虑的翻译假设（即候选译文）的数量。翻译假设产生于训练过程和系统的实际翻译过程。
- “训练时长”（Duration）是指训练轮次的数量。一“轮”(epoch) 指完整训练一次训练集的所有句子。每一轮都包含历遍全部数据所需的所有训练步骤。[14]
- “验证频率”(Validation frequency) 指每次评估训练情况前的训练步骤数量。通常来看，每一轮训练都会进行多次验证。
- “停止条件”(Stopping condition) 规定在机翻引擎性能没有改善的情况下，所允许的最大验证执行次数。若达到预设的最大验证次数，无论轮次初始设置为多少，系统都会停止训练。训练效果可通过任何单个或多个自动评测指标来衡量。常见的自动评测指标有 BLEU、chrF1 和困惑度 (perplexity)。[15]

这些参数通常都设置了默认值，这些默认值是开发者在为特定环境优化训练过程后设置的。

设置好参数或采用默认参数后，训练流程如下：每次迭代都将一批训练数据输入神经网络，计算该批数据每个句子的输出译文，计算误差损失，更新权重，然后再重复上述过程处理下一批数据！完成预先设定的迭代数（“验证频率”参数设置）后，评测机翻引擎性能，然后继续训练。若更多的训练未能改善引擎性能或性能变差，则训练停止。模型训练完成！

[13]近几年，以“分词单位”（参见Rossi & Carré 2025 [本卷]）代替句子来计算批量大小已成为最常用的批量类型，以使不同批量的大小更接近。

[14]考虑到神经网络机器翻译需使用大量数据，通过轮次来衡量训练时长可能不切实际。除了轮次，还可以使用与特定批量大小有关的步骤数，来单词计算模型、语言对或数据量来的训练时长。

[15]在自然语言处理中，具体到机器翻译，困惑度用于衡量翻译模型在翻译时预测下一个单词的不确定程度。若翻译模型为给定目标语句子中的每个单词/分词单位赋予高概率值，则为低困惑度。欲进一步了解 BLEU 和 chrF1，请参见Rossi & Carré (2025 [本卷])。

### 4.6 测试定制化模型

机器翻译系统训练结束后，需要对其进行测试或评测。测试方法总结如下：

- 自行（或请别人）测试！如果你会说正在研究的语言，了解训练系统的用途，且有足够的时间，不妨用系统来翻译一些句子，看看质量如何！使用正确的工具，不仅可以查看到系统的最佳译文，还可以得到它为每个句子提供的 $n$ 个候选译文列表。

- 评测质量！自动评测指标可以计算系统的性能得分（参见Rossi & Carré 2025 [本卷]）。大多数指标都采用对比系统译文与专业译者的“参考”译文的方法。这些指标的意义有时差别较大：有的虽然适用于比较两个不同的系统，但数字本身没有太大意义，例如基于 $n$-gram 或字符的评测指标，包括 BLEU(Papineni et al. 2002)、METEOR(Denkowski & Lavie 2014) 和 CHRF1-3(Popović 2015)。还有的指标有助于测量将机器译文变为专业译文所需的工作量，如 WER(Popović & Ney 2007) 或 TER(Snover et al. 2006) 等用来测量译后编辑工作量的指标。此外，还有一些指标能够体现机器译文的特点。文本层面的测量指标涉及词汇的多样性或密度。最后，有些指标用于对系统进行排名，指出你的系统是否优于其他系统。

- 从实际使用中获得反馈！根据系统用途，你可以评估该系统在真实的专业或日常使用场景中是否有用。你训练系统是为了帮人写电子邮件吗？是的话，那就让人用它来写邮件，然后向你反馈使用体验。你训练系统是为了帮人看懂食谱吗？是的话，那就让人按照系统翻译的食谱做菜，然后给你反馈。你训练系统是为了翻译法律文件吗？是的话，那就让人用来翻译，然后跟你分享使用心得，如系统是否能节省翻译时间等。通过收集这些反馈，你不仅能了解当前的系统状况，还能提供改进系统的思路。

最后，历尽艰辛训练出定制化机器翻译系统，你可能想通过前文所述的测试方法，来对比自己的定制化系统和通用系统的译文。如果前者优于后者，那就算成功了！否则，再接再厉吧！

## 5 结语

本章概述了如何定制开发机器翻译系统。作者首先区分定制化和通用机器翻译系统，强调了管理定制预期的重要性，然后介绍了定制开发神经网络机器翻译系统所涉及的专业角色，并思考机器翻译如何融入翻译工作流程。此外，本章还讨论了基于数据和技术的定制方法，建议将其与现实生活中的学

习过程进行类比。最后，作者从实践角度出发，介绍了相关工具、定制策略、数据搜集和准备、系统训练和结果测试，以帮助读者获得机器翻译系统定制化的实践经验。

## References

Denkowski, Michael & Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on Statistical Machine Translation*, 376–380. Baltimore: Association for Computational Linguistics. DOI: 10.3115/v1/W14-3348.

Kenny, Dorothy. 2025. 人工翻译和机器翻译. In Dorothy Kenny (ed.), 机器翻译知识普及: 为人工智能时代的用户赋能, 19–38. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922287.

Khayrallah, Huda & Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd workshop on neural machine translation and generation*, 74–83. Melbourne: Association for Computational Linguistics. https://aclanthology.org/W18-2709.pdf.

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge: Cambridge University Press.

Mohiuddin, Tasnim, Philipp Koehn, Vishrav Chaudhary, James Cross, Shruti Bhosale & Shafiq Joty. 2022. *Data selection curriculum for neural machine translation*. DOI: 10.48550/ARXIV.2203.13867.

O'Brien, Sharon. 2025. 如何处理机器翻译的错误：译后编辑. In Dorothy Kenny (ed.), 机器翻译知识普及: 为人工智能时代的用户赋能, 83–94. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922293.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.

Pérez-Ortiz, Juan Antonio, Mikel L. Forcada & Felipe Sánchez-Martínez. 2025. 神经机器翻译的工作原理. In Dorothy Kenny (ed.), 机器翻译知识普及: 为人工智能时代的用户赋能, 111–128. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922297.

Popović, Maja. 2015. Chrf: Character n-gram f-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, 392–395. Association for Computational Linguistics. 10.18653/v1/W15-3049.

Popović, Maja & Hermann Ney. 2007. Word error rates. In *Proceedings of the Second Workshop on Statistical Machine Translation* (StatMT '07), 48–55. Prague, Czech Republic. DOI: 10.3115/1626355.1626362.

Rossi, Caroline & Alice Carré. 2025. 如何选择合适的神经机器翻译解决方案: 机器翻译质量评测. In Dorothy Kenny (ed.), 机器翻译知识普及: 为人工智能时代的用户赋能, 39–63. Berlin: Language Science Press. DOI: 10.5281/zenodo.14922289.

Saunders, Danielle. 2021. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *CoRR* abs/2104.06951. https://arxiv.org/abs/2104.06951.

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th conference of the Association for Machine Translation in the Americas: Technical papers*, 223–231. Cambridge, Massachusetts: Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.25/.