第 1 章　自然语言处理概要

自然语言处理（Natural Language Processing，NLP）既是人工智能（Artificial Intelligence，AI）的一个分支，也是计算机科学（computer science）和语言学（linguistics）的交叉学科，它的目标是运用计算机处理、理解自然语言，从而完成一些有意义的信息处理任务。作为交叉学科，自然语言处理又称为计算语言学（computational linguistics）。自然语言处理是人工智能的挑战性分支，从信息感知角度来说，有别于其他分支，如计算机图像/视觉（image/vision）、语音（speech/voice）处理，自然语言处理的目标是文本（text）。这里需要注意的是，虽然“自然语言处理”这一术语中带有“语言”二字，但其更多指向的是文本对象；而理论语言学中的“语言”多指语音对象[①]。这是因为，最初不管是自然语言处理还是计算语言学，是既包含文本处理也包含语音处理的，直到 20 世纪 90 年代，自然语言处理和语音处理才在学术界分家。理论语言学（为区分于计算语言学，也称之为纯语言学）的研究具有数百年的悠久历史；相比之下，自然语言处理所属的现代计算机学科的存在还不到 80 年，但是发展迅猛，研究子领域之间也分分合合，术语含义也不断迁移。

自然语言（有时也称为人类语言）的出现是智能展示的相对完备形式。从进化尺度上看，人类进化史约上百万年，人类语言（不要求有文字）的出现应不早于十万年前，而文字（现代自然语言处理的研究对象）的出现更是在一万年以内，也就是说自然语言处理研究的是全部人类历史最近 1%时间内的产物。

在进入自然语言处理学习之前，推荐读者先修以下课程：

（1）数学基础课程，包括数学分析、概率论与数理统计、线性代数和矩阵理论、解析几何等。

（2）计算机基础课程，包括数据结构与算法基础、编程语言（C/C++、Python 等）以及机器学习基础。

在机器学习的实际操作上，特别建议读者最好提前熟悉 PyTorch、TensorFlow 等深度学习工具。

本章将介绍自然语言处理的基本背景，其中包括：①概念和术语；②技术性挑战以及机器翻译的背景介绍；③语言处理层次的概念；④结合自然语言处理应用介绍其历史发展；⑤自然语言处理相关的学术出版体系。

1.1　自然语言处理的概念和术语

1.1.1　自然语言

自然语言指人类语言，比如汉语、英语、德语或法语。听、说、读、写是自然语言

① 当然这不意味着理论语言学只研究语音所指的语言对象。理论语言学也研究文本，如乔姆斯基的句法学、韩礼德的功能语法都是研究文本上的句法。在索绪尔的共时研究中，文本也是重要的对象。在理论语言学中，专门研究语音的是音位学和音系学。

最常见的运用方式。人类语言也称作自然（nature）语言，是因为它是自然进化的产物，是生产生活斗争中随着需求变化而产生的。作为比较，计算机编程语言是由人类创造的一套符号和规则系统，用来将一套指令完整、精确地传达给一台计算机。编程语言在编写时带有一定的意图（语义），同时遵循关于变量、函数、不同类型的括号等的规则（语法）。编程语言与自然语言有很大区别。首先，相比于自然语言中使用的词典、词义、语法规则，编程语言中用到的关键词很少并且含义确定，涉及的语法更少也更简单。其次，编程语言所有的规则和定义都是预先设计好的，这使得它们能够被完整、精确地描述和研究，不会造成任何疑惑；而自然语言由于多义词、同义词的存在，很容易产生歧义现象。最后，由于编程语言中规则的严格性，使用者不能随意发挥和篡改；而自然语言中充满了不规范、不完整甚至错误的表达，例如方言、俚语、行话、拼写错误、不规则的标点符号等。正是自然语言的这些特点增加了自然语言处理的难度，也使其远比利用编译器处理编程语言更复杂。

某种意义上，自然语言处理针对的对象——人类语言是符号系统中最复杂的，因此，自然语言处理领域发展了最为精巧的符号系统处理技术。幸运的是，针对符号系统的信息处理并不都是如此具有挑战性，编程语言就是这样一个处理技术“友好”的符号系统。在最近 20 年的信息处理实践中，自然语言处理技术不仅用于自然语言，也被广泛迁移到软件信息工程、编程语言处理、生物信息学以及化学信息学等领域，针对的处理对象（符号系统）涵盖编程语言、蛋白质序列、人类基因组序列以及化学分子式表示（如 SMILES 码）等。当自然语言处理技术针对的对象不限于自然语言，甚至开始跨界处理多种不同类型的符号系统时，自然语言处理就正在现实中走向某种意义上的“广义符号处理”。

自然语言有两种形式：书面形式和口语形式，分别对应现在的自然语言处理和语音处理。在 20 世纪 90 年代之前，自然语言处理主要依赖于规则方法，而非今天流行的以统计为基础的方法。虽然实际能应用的统计和数学知识甚至已经存在了一两个世纪，但是由于当时的计算能力极为有限，计算机硬件条件无法支持需消耗巨大计算资源的统计方法，从而使得历史上的自然语言处理界只能囿于规则方法。自然语言处理是众多子任务的总和，其中也曾经包含语音处理任务。语音处理可以简单分为语音合成和语音识别两个处理方向相反的子任务，相对其他繁杂的文本处理来说，其任务模式较为单一，同时又非常依赖于统计语言模型，因此，开始就需要统计方法支持的语音处理与一直囿于规则方法的早期自然语言处理的其他分支早早就分道扬镳了。到了 21 世纪，由于大众可用的计算能力的普遍提升，统计方法才开始被越来越广泛地应用于自然语言处理的各个分支。由于这样的历史，导致了现在的“自然语言处理”中的“语言”颇为吊诡地仅仅指的是文本。

1.1.2 自然语言处理与自然语言理解

自然语言处理集合了通过算法、统计或常识等处理语言的各类方法。自然语言处理任务大致可以分为两类：

（1）基本任务，包括语言建模和表示以及语言结构和分析，后者又包括形态分析（含

分词）、句法分析、语义分析和篇章分析等。

（2）应用型任务，包括对话系统、机器翻译、语言理解和语言推理等任务。

很长一段时间以来，作为人工智能的分支，业界更多地用“自然语言理解”称呼“自然语言处理”这一研究方向。自然语言理解（Natural Language Understanding，NLU）研究的是对某种自然语言文本的真正理解，被认为是人工智能的核心难题，甚至是终极难题。实际上，“自然语言处理”这一术语的广泛使用大约从 20 世纪 90 年代才开始，比“自然语言理解”这一术语的使用晚很多。21 世纪以来的很一段时间内，这两个术语就所指的研究方向和研究内容而言都是一致的。两个术语的选用只是研究者的偏好问题。“自然语言理解”这一术语更多地被人工智能中从事交叉方向研究（例如计算机视觉、非单调推理、机器学习等）的研究者所采用；而“自然语言处理”这一术语则越来越多地被从事单一语言处理研究的研究者所采用。

自然语言理解一度也被认为是自然语言处理的下一个预期的阶段，但在很长一段时间内，由于其过大的挑战性，“自然语言理解”只是人们憧憬的美好目标，而无具体的实际操作实践。这也导致了用“自然语言理解”指代整个研究领域的这一做法的减少，相关研究者越来越愿意称自己研究的是“自然语言处理”。随着可用信息资源的增多，使得更复杂的神经网络成为可能，研究者向着让计算机能够真正理解人类语言的前沿迈进。现在“自然语言理解”的使用在减少，并且“自然语言理解”的概念被逐渐窄化，开始限定于指真正、具体的、在 20 世纪第二个 10 年才首次出现的自然语言理解任务，如现在流行的机器阅读理解（Machine Reading Comprehension，MRC）和自然语言推理（Natural Language Inference，NLI）这两个任务。今天，针对这两个术语的使用，研究者回归了自然语言理解和自然语言处理字面意义所指的常态，达成了相对一致的共识：“自然语言处理”是方法和手段，而“自然语言理解”是目标。

事实上，自然语言理解的挑战性很多就是源于自然语言处理上的挑战。例如，词、短语组合的多样性会导致不同的语言含义，不联系上下文以及环境约束会造成语言歧义性，语言作为开放符号集合可以任意地发明创造一些新的表达方式，语言理解多需要外部知识支撑，等等，导致了机器在自然语言理解上的表现至今还远不如人类。另外，自然语言理解的任务目标除了要求计算机理解语言的字面含义（语义）以外，还需要理解语言的“言外之意”（语用），例如“我觉得好冷啊”在语义上表明冷，而语用上就需要探究这句话背后的含义，例如说话者是否想要调高空调温度等。在人机对话中，自然语言交互涉及语法、语义、语用 3 个层面，如果希望机器能够真正读懂人类语言的复杂语义，研究者需要在自然语言处理的基础上综合引入认知语言学、心理语言学、社会语言学等学科的知识信息，在语义理解的基础上增加意图识别和情感判断，以弥补纯粹的语言处理的不足，让计算机能够真正读懂人类语言的复杂语义以及背后的意图和情感，并在此基础上给予对话者拟人的反馈，从而达到更好的交互效果。

1.1.3　计算语言学

不同于单语言专业（例如中文系、英文系），理论语言学是研究多种人类语言共性的学科，至少已经有两百年的历史。在传统视角上，计算语言学是以理论语言学为主而

发展的交叉学科，建立该学科最初的目的是提出一种可被计算机处理的语言学理论、框架、模型。而后来的计算语言学则研究计算机在语言研究中的应用，或使用计算机研究语言学。计算语言学在计算机于 20 世纪 40 年代问世后不久就开始出现，比“自然语言处理”这一术语的产生早得多。计算语言学也可以解释为从理论语言学的角度看待自然语言处理这个方向。

从计算机科学的角度来看，自然语言处理当然是计算机科学的一个子集，特别是其中人工智能方向的一个分支方向，其最终目的是让计算机能够理解自然语言。在科学史上，从语言学专业和计算机专业出发分别进行自然语言处理研究，形成了异曲同工的有趣关系。这也可以从自然语言处理领域两个非常重要的会议——ACL（Association for Computational Linguistics，计算语言学协会，由语言学家创办）会议和 EMNLP（Empirical Methods in Natural Language Processing，自然语言处理的经验方法，由计算机专业人员创办）会议的创办者来源可以看出。现在，计算语言学与自然语言处理的研究内容已经相似到难以区分，两者的界限逐渐模糊，在指代领域名称的术语运用上，“计算语言学”已经被认为等同于“自然语言处理”。

计算语言学（或者自然语言处理）作为一门交叉学科的特性是非常明显的，图 1.1展示了计算语言学/自然语言处理与相关领域的关系。计算语言学/自然语言处理的方法、目标是工程的，语言是智能的关键特性，因此，它显然是人工智能中的一个关键分支。首先，计算语言学/自然语言处理当然涉及计算机科学，因为其中包含了十分严格的算法与数学基础；其次，计算语言学/自然语言处理还涉及认知科学，有别于动物大脑，语言处理是人脑的一个特有功能；最后，计算语言学/自然语言处理也涉及生理学、心理学、哲学、语言学等领域。因此，计算语言学/自然语言处理不是计算机科学 + 语言学的简单交叉，而是一个涵盖极广、多学科复杂交叉的研究方向和学科。

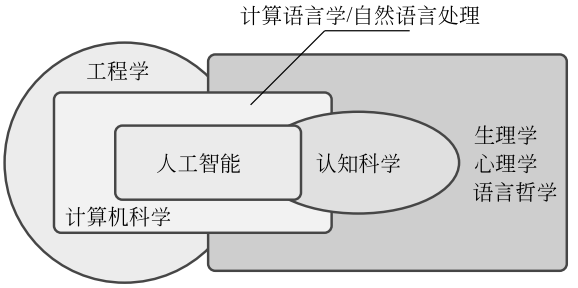


图 1.1 计算语言学/自然语言处理与相关领域的关系

1.2 自然语言处理的技术性挑战

现代人工智能所依赖的数学工具、算法等基础在几十年前就已经完成，今天的人工智能系统不过是在大众能够承担得起所需要的计算能力时，将这些技术实现并适时推送给最终用户。

知识已被公认为是人工智能的核心主题，常识知识又是普遍人工智能必须面对的终极挑战。在所有人工智能分支中，或许只有自然语言处理被迫需要承担两类知识——常识知识与语言学知识的处理和解析任务。后者属于自然语言处理这一领域独一无二的需求。可以这么对比：计算机视觉的处理并不需要依赖一个可观的“视觉学知识”才能达成。自然语言处理作为一门交叉学科具有更大的相对独立性，而它的交叉来源之一——理论语言学有着比现代计算机科学长得多的悠久学术传统。正是因为在人工智能的常识知识困境之外，自然语言处理还需要谨慎处理语言学知识这一瓶颈，才使得这一研究方向在人工智能各个分支之中尤其具有挑战性。

从词汇层面看，人类语言内的上述两类知识并不难以区别。自然语言中的常识知识以命名实体（named entity）以及实体之间的关系展现，前者如“新型冠状病毒肺炎”“苏格兰场”“上海交通大学”等，后者如“上海”与“中国”的包含关系以及“古特雷斯”和“联合国秘书长”的同位关系等；语言学知识相对抽象，直观上可以理解为人类语言内除了常识知识之外所有有关语言本身形式化、构成规则的知识线索，例如词性（part-of-speech）、句法（syntax）、形式语义（formal semantics）等。从词汇层面看，如果一个词非命名实体词，则它体现形式语义的时候展示的就是语言学知识。

目前为止的自然语言处理的技术实践很大程度上其实是在运用语言学知识解析这两类知识。在一般的人工智能分支之中，只处理一类知识（常识知识或很窄的领域知识）尚且十分困难，更不用说，自然语言处理的任务需要同时面对两类知识解构的挑战。由此可以看出自然语言处理的技术挑战性非同一般，同时需要处理两类知识成为自然语言处理挑战性的根本来源。

从具体的语言学知识形式上说，自然语言处理的挑战大体包括以下几方面：

（1）歧义（ambiguity）问题。相比于精确、唯一、无歧义定义的计算机编程语言，自然语言的表达形式和语义之间的映射有一对一、多对一、一对多或多对多 4 种类型。例如，英语表达之中的一对多映射“turn right”与“that's right”中“right”就具有歧义性。一对多映射一般情况下需要专门输入额外的大量领域知识，才能在目标形式表示中做出正确的解析选择。

（2）知识依赖问题。

① 修饰语附着（modifier attachment）问题。例如，英语句子“Give me all the employees in a division making more than $50 000”中并没有说清楚“making more than $50 000”所修饰的是“employees”还是“division”。此类连续修饰语的附着问题根源在于语言表达形式的线性特性与其含义之间的非线性特性之间的本质冲突。具体来说，语言在书写或者交流的时候必然是线性的，即，文字需要从左到右（或从右到左）书写，口语需要逐个吐出一个个词语，但这种客观受限的线性表达事实上无法精确展现修饰关系的非线性结构。例如，图 1.2展示的修饰结构实际上的语序可能是 $OM_1M_4M_2M_3$，而如果给定这样一个语序，在没有额外的信息线索支持下，无法确切还原出图 1.2中的这种修饰关系。需要指出的是，修饰语附着问题并非由于一个语言支持前置修饰语语法或后置修饰语语法导致的，也不是因为同时支持前后置修饰语导致的。如果说英语支持后置修饰语（上面的例句即是）导致了此类问题，那就无法解释中文这种仅支持前置修饰语的

语言也会在连续的修饰语结构中出现修饰语附着的消歧问题。

由于表达上的线性手段无法覆盖内在语义结构非线性的本质困难，修饰语附着问题不太可能通过重新定义语言学规则加以解决，也不太可能仅利用语言学知识精确有效求解。首先，如果规定修饰规则，则会使语言使用的自由度大大受限而导致不便。事实上，在语言的实际使用中定义规则是相当不现实的。其次，自然语言之所以是自然语言，是由于其语言表达的线性模式也不太可能改变，同时也不可能在表达时为每句话都画出类似图 1.2的非线性修饰结构。最后，仅利用语言学知识，例如词性、语法、形式语义等，均无法一般性地有效解决修饰语附着问题。例如，已知“making more than \$50 000”、“employees”和“division”是 3 个句法成分对于求解这个问题没有太大帮助。就人类经验而言，实际上需要具备常识知识，比如“employee”和“division”正常情况下会赚多少钱来判断“making more than \$50 000”具体修饰的对象。如果训练一个统计机器学习模型决定修饰对象以求解修饰语附着问题，则该模型会简单地向训练损失最小的方向做出猜测，只要没有合理引入这里所需的常识知识，这种方法实际上还是忽视了认知上的根本理由，而仅仅是对于统计的规则化实现。

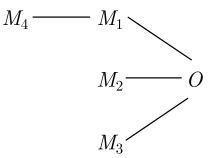


图 1.2 修饰结构：M_4 修饰 M_1，M_1、M_2、M_3 修饰 O

② 量词范围（quantifier scoping）问题。以英语为例，在逻辑上，某些限定语，如“the”、“each”或“what”，表示“通用（所有）”（∀）或“存在”（∃），对它们所指管辖范围可能会有多种理解。此类问题在求解形式上和修饰语附着问题类似，都是要求判断句子之中的两个词或成分之间有无依赖关系。

（3）省略表达（elliptical utterances）问题。在人类语言中，语言成分的省略是一个普遍现象，或是出于特定的语法设计（如日语在实际使用中会系统性地省略各类语法成分），或是在复杂的上下文交互中为了略去对话双方已知的内容而寻求高效的沟通。然而，精确的语言处理和语言结构分析需要以完整的上下文为前提才能进行，因此，相应的语言处理任务的基本需求是从省略的表达中恢复出完整的非省略的上下文，以供后续进一步处理。以对话为例，一个简化或省略的问句的解释要取决于先前的问句及其解释。例如，询问“Who is the manager of the automobile division?”，然后接着问“of aircraft?”，这时就需要结合上文才能恢复完整形式，确定是在询问“经理是谁”。如果这里是多轮对话，完整信息可能在更早轮的语句中才能获得，那么这个问题会变得更加困难。

如我们所知，无论是在空间上还是在时间上，人类语言都是一个开放的符号集合，在历史长河之中，不断有新的语言表达形式被发明创造出来，语言的表达形式和内涵之间的关联不断变迁。语言的这些时空演化的动态性进一步给要求精确、高效的自然语言处理增加了难度。

1.3　机器翻译

机器翻译（Machine Translation，MT）研究如何利用计算机自动实现不同语言之间的相互转换，是自然语言处理的重要研究领域。机器翻译是一个重要的、有着重大需求和实际运用场景的语言处理应用，也是易于理解的多语种语言处理的典型案例。早在 1949 年，沃伦 • 韦弗（Warren Weaver）即提出计算机可能对“解决世界范围的翻译问题”有用[1]，思路就是针对翻译的源语言和目标语言构造双语词典，进行转换后再重新组合。经过 50 多年的努力，尽管取得了巨大成就，21 世纪初的机器翻译系统仍然只能产生质量一般的结果，仅适用于粗略了解外文大意的场景。在 21 世纪的前 20 年结束的时候，机器翻译研究取得了更为显著的成就和进展，然而依然远远不适用于产出正式文档。

在早期的机器翻译实践中，或者受制于有限的计算资源，或者由于认识不足，人们一度以为可以通过枚举所有的翻译规则的方式建造实用的机器翻译系统，这一思路一直到 20 世纪 90 年代统计机器翻译方法崛起时还在被人尝试。在类似情感分析这样简单的语言处理任务中，的确可以相对轻松地列出所有正面词和负面词的列表，并构造出正面词和负面词的转换规则，依靠这些手段就能建立一个可工作的情感分析系统。但是，对于机器翻译来说，这种方法最终被证明代价过于高昂，因为没有人能够完备地枚举出从一门语言翻译到另一门语言的所有规则。通过长期实践，研究者已经意识到人类语言翻译是一种复杂的认知和处理能力，涉及不同类型的知识：

（1）句子结构。不同的语言遵循不同的句子结构。例如，中文和英文都遵循“主语—谓语—宾语”的形式，日语和印地语遵循“主语—宾语—谓语”的形式，而阿拉伯语则遵循“谓语—宾语—主语”的形式。如果这些语言属于同一语系，语法差异或许会相对较小，例如，英语和德语属于印欧语系（Indo-European family），泰米尔语和泰卢固语属于达罗毗荼语系（Dravidian family），等等。具有巨大句子结构差异的语言对之间的翻译会比具有相同或类似句子结构的语言对之间的翻译困难得多。

（2）词义。一个词可能会有很多种不同的含义。例如，英语句子“Please book my ticket for tomorrow”与“Please buy that book for me”中的“book”，前者指“预订”，后者指“书”。人可以从词语的上下文中理解其含义，但是这对于计算机而言是困难的。针对双语处理的机器翻译会面临进一步的困难，因为翻译的源语言和目标语言之间的词义歧义方式会完全不同。例如，英语的“book”兼具“预订”和“书”的义项；而中文的“书”并无“预订”这一含义，但是有英语的“book”所不具备的“书写”这一含义。

（3）常识。即关于世界的广泛共享信息。人类在对自然语言中的信息进行分析时，在一定程度上依赖于一些第三方信息，这也是语言学家耶霍舒亚 • 巴希勒（Yehoshua Bar-Hillel）宣称机器翻译不可能实现的理由，他举的例子被称为巴希勒悖论（Bar-Hillel paradox）[2]：“The pen is in the box”与“The box is in the pen”，前者的“pen”翻译成钢笔，而后者的“pen”翻译成围栏。要想得到正确的翻译，一种方法是根据上下文推理，但是在没有上下文的情况下，大多数人也能够正确翻译这两个句子，因为他们知道“pen”（钢笔）比“box”（盒子）小，“box”（盒子）比“pen”（围栏）小，并且只有

较小的东西才能放在较大东西的里面。而机器要进行正确的翻译也需要具备这些额外的常识。

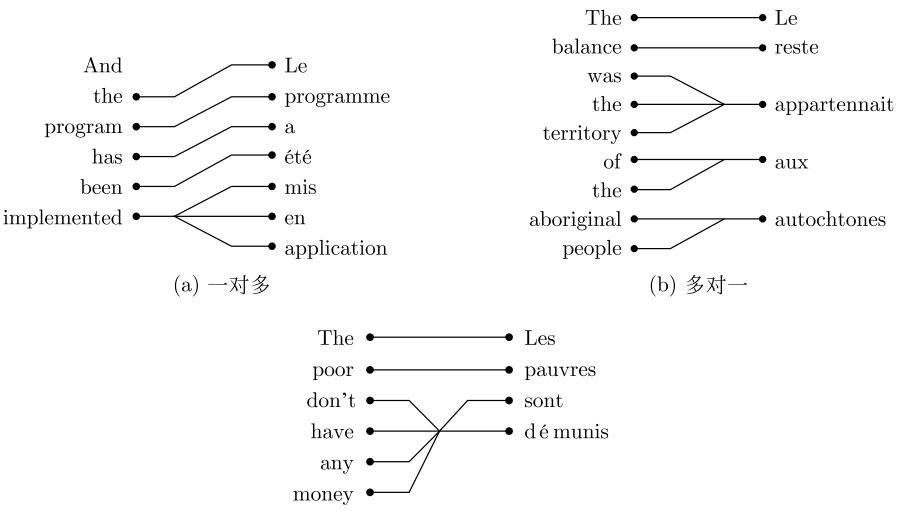
除此以外，机器翻译还涉及听众模型（用户模型）、对话规则（对话翻译）等方面的知识。这些要素实际上已经涉及自然语言处理、自然语言理解中几乎所有内容要素，这些要素组合在一起已经被证明是一个非常复杂的任务。1964 年，约翰·罗宾森·皮尔斯（John Robinson Pierce）发表了自动语言处理咨询委员会（Automatic Language Processing Advisory Committee，ALPAC）报告[①]，否定了短期内机器翻译研究能产生有意义影响力的可能性。从此，机器翻译进入了长达 30 年的低谷期。

20 世纪 80、90 年代之交，在 IBM 研究中心超级计算机的算力支持下，IBM 的研究者提出了现在称之为 IBM 模型的翻译对齐学习模型，从而开启了统计机器翻译（Statistical Machine Translation，SMT）[3,4] 的时代，机器翻译也从低谷期开始复苏。21 世纪初，统计机器翻译的另外两个关键要素也得以有效建立：最小错误率训练（Minimum Error Rate Training，MERT）[5] 方法提出，用区分式机器学习方法自动集成 IBM 模型和 n 元语言模型，帮助生成稳定的翻译文本；翻译质量自动得分评估方法——BLEU [6] 也被提出并被广泛接受，结束了对翻译质量评估方法的争议，大大缓解了根据开发集调参时需要人为干预进行翻译质量评估的不便，使得机器翻译模型的全自动优化成为可能。所有这些进展都推动统计机器翻译进入全盛时期。至此，IBM 模型、MERT 方法以及 BLEU 评估方法成为统计机器翻译的三大技术支柱。

机器翻译需要双语平行语料库作为训练集，其中的句子或段落会以某一种语言表述并且对应到另一种语言表述的相应句子或段落。在传统的统计机器学习中，这些翻译系统非常复杂，一般被分为几个子模块，如翻译模型、语言模型、调序模型等，这些模型相互独立，以管线方式组合在一起，分别进行优化。翻译模型需要将源语言与目标语言的词对齐（alignment），即确定源语言中的哪些词语对应目标语言的哪些词语。对齐是机器翻译中的关键难题，图 1.3展示了英语和法语句子之间的对齐示例。一个双语句对之间的词对齐模式有一对多、多对一、多对多等；甚至有的源语言中的词不用被翻译，因而也无须对齐。所有对齐的可能性数量庞大到组合爆炸，这使得对齐学习问题变得非常困难。在利用对齐模型获得了所有潜在对齐之后，对于源语言句子之中的每个词或者每一个短语都会有大量的翻译候选，整个源语言句子的翻译结果存在于这些对齐候选组合形成的一个巨大的搜索空间之中。在统计机器翻译的解码过程中，通常用一个 n 元语言模型确定哪一个翻译组合更好，以决定最优的翻译结果。

2014 年，Google DeepMind 提出的神经机器翻译（Neural Machine Translation，NMT）模型[7,8] 使得机器翻译进入了新的时代。神经机器翻译模型抛弃了 IBM 模型等组件以及 MERT 训练方式，仍然使用 BLEU 作为自动评估标准。相比于传统的统计机器翻译，神经机器翻译利用具备表示学习机制的深度神经网络对整个翻译过程建模，以一种端到端（end-to-end）的方式对这个网络进行一次性训练优化，只需要关注目标函数，整个翻译过程都能在一个模型中同步学习。相比于统计机器翻译，即使在有图形处理器（Graphics Processing Unit，GPU）加速的情形下，神经机器翻译也需要使用更多

① http://www.hutchinsweb.me.uk/MTNI-14-1996.pdf.



(a) 一对多　(b) 多对一

(c) 多对多

图 1.3　英语和法语句子之间的对齐示例

的计算资源以及更长的训练时间。目前为止，已经证明神经机器翻译更有效，是比统计机器翻译更好的建模方式，前者可以取代后者。但神经机器翻译仍存在一些悬而未决的问题，例如难以利用先验知识和约束机制，过度翻译和翻译不充分，训练速度慢，处理生僻词效率低，甚至有时漏译原句中的词，等等[9]。正是由于人类语言的巨大开放性和无比的灵活性，机器翻译，包括最新进展下的神经机器翻译，依然面临很多挑战。

1.4　语言处理层次

自然语言处理的研究者期望他们的工作能够在更好地理解语言、更智能的本质基础上有助于开发实用、有效的语言处理和语言理解系统。人类语言是一个复杂的符号系统，显然，一蹴而就的处理和理解不太现实，在理论语言学悠久的研究传统和被广泛接受的形式化理论成果的启发下，詹姆斯 • 艾伦（James Allen）教授提出了语言分析的六大层次[10]：

（1）形态分析（morphological analysis）。

（2）句法分析（syntactic analysis）。

（3）语义分析（semantic analysis）。

（4）语用分析（pragmatic analysis）。

（5）篇章分析（discourse analysis）。

（6）世界知识分析（common sense analysis）。

尽管不是每位自然语言处理的研究者都愿意接受这样的分类分层体系，但目前为止的自然语言处理的各个基本任务大体上都能被这样定义的 6 层处理分析所覆盖，因此，在实际工作中研究者们或多或少、有意无意地遵循了这样的处理思路和体系，设计和发展相应的数据集、任务定义。

1. 形态分析

大多数自然语言分析系统通常首先需要将文本分割为有语言学意义的符号单元，这产生了形态分析任务的需求，而形态分析是由书写形式决定的。纯语言学定义的形态主要指词（默认的语言基本单元）的书写规则。形态分析指的是从完整书写的文本之中识别、提取语言基本单元。这里的完整书写的文本可以指句子（如中文），也可以指单个词（如英文以及大多数字母拼写的文字），这里的语言基本单元通常指的是词（如中文），也指词的原形、词根（如英文）。形态分析的需求从根本上来说是因为相应的文字系统对于精确处理不友好而导致的。大多数人类语言遵循字、词、句的层次堆叠的组织形式。词作为一般语言处理的基石，是绝大部分语言处理系统要优先识别的对象，然而，很多语言所使用的文字系统和文字书写方式给确切的语言处理创造了先天性的障碍，大量的语言在书写形式上并未提供可靠、精确的词形式。由于书写形式和语法组织差异，不同语言的形态分析具有不同的内涵，其区分标准在于语言的形态是否丰富，即词的构成规则体系是否足够复杂。对于形态简单乃至无形态的语言和形态丰富的语言而言，其形态分析任务的定义是完全不同的。

以中文为例，中文首先是一个形态简单到可以认为无形态的语言，其书写以汉字这一大字符集为基础，句子的整体书写字词不分，词与词之间并未有类似空格这样明确的标识。现代汉语早已不是一字一词的古代汉语，据大范围的词频统计（如表 1.1所示），约 47%的词是单字词，约 45%的词是双字词，剩下是更多字构成的多字词。因此，中文的形态分析任务的目标是从连续书写的句子之中切分出词，这一任务也被称为中文分词（Chinese word segmentation）。分词这一操作需求在所有词的界限不明确的语言文本处理中都会存在，而不仅限于中文，也不仅限于非拼音文字。实际上主要东亚语言，如日文、韩文/朝鲜文、越南文、缅甸文、泰文，都有分词的需求。日文和韩文曾经或继续使用汉字作为书写符号。而现代越南文已经成为以拉丁字母为基础的文字，同样有分词的要求，原因在于越南文书写时以音节为单位，所有音节之间都有空格（而中文句子中所有字之间都没有空格），词的界限依然没有明确标识[①]。

表 1.1 SIGHAN Bakeoff-2（2005）提供的 4 种切分数据集的词长分布（%）[11]

	数据集及大小（词数）			
词长（字数）	AS(5.45M)	CityU(1.46M)	MSRA(2.37M)	PKU(1.1M)
1	57.1	46.9	47.2	47.5
2	37.9	45.5	43.9	45.0
3	3.6	1.3	4.8	5.0
4	1.0	0.2	2.4	2.1
5	0.2	0.2	0.9	0.6
⩽5	99.7	99.9	99.0	99.8
⩾6	0.3	0.1	1.0	0.2

① 现代越南文之所以采取这样的书写形式是因为古典越南文的书写是基于汉字的扩展字符集，称之为字喃。当一字一音一义的字喃被转为拼音之后，越南文内部固有的词界限不清晰，只能通过书写的音节（一个音节代表一个字喃）之间普遍存在的空格表达。