自然语言处理（Natural Language Processing，**NLP**）是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。

NLP 难点：主要包括语义的模糊性、语言表达的多样性，语言的一词多义，语言的知识依赖，语言的上下文依赖，还有不同语言中的特有问题，如英文中代词的指代关系，中文中的儿化音、多音字等等，以下是一些举例：

语义的模糊性。对于图像而言，每个像素点都有一个数值，而一个数值就代表一种特定的颜色，这种编码形式是确定的，比如一串黑色的点拼起来，就代表着一条直线。而语言的表述中往往存在背后的含义，即语义。如下所示，语义的表现是模糊的、因人而异的，因此成为了 NLP 的一大难点。

表达	含义
我好饿	想要吃东西
NLP 好有趣	想要继续学习

语言表达的多样性。NLP 中语言的表达是丰富的，如下所示，可以委婉或直接、抒情或亢奋，在处理过程中我们希望保证同一种语义的不同表现形式尽量具有相同或相似的编码，这也为词义、句意的刻画过程增加了难度。

同一含义的多种表达
我们新推出了大模型课程。
我们上线了大模型课程。
大模型课程是我们公司的新产品。

语言的一词多义 。每种语言都或多或少存在一词多义的情况，且有时一个词的几个含义差别还会很大，如下所示。那么在编码过程中，就很难在某一特征空间内直接采用文字与数值的一一对应关系，也为文本的表示增加了难度。

例句	一词多义
中关村的苹果不错	水果 or 品牌
你这人真有意思，你要是这意思的话那就太不够意思了	有趣 or 含义 or 义气

语言的知识依赖。我们在阅读专业领域知识的时候，也会碰到如下所示的情况，明明每个字都认识，每个字的含义我们也清楚，但把它们连起来就读不懂了。同样，机器在理解自然语言时也会碰到专业领域的训练语料少，领域内简称、缩写名词繁多等难题。

例句	专有名词
通过原位原子力显微镜纳米力谱和透射电子显微镜能量色散 X 射线光谱等表征技术，研究人员成功实现了锡化合物纳米结构的无损测量及原位变形行为的观测。	原位原子力显微镜
	纳米力谱
	透射电子显微镜
	能量色散 X 射线光谱

语言的上下文依赖。很多时候我们在理解文中某一些词时，需要依赖上下文的内容进行推断，或者甚至需要通过整篇文章的主题来推断。如下所示，单词的含义取决于不同领域的上下文，只有让计算机前后关联地去理解每一个词，才能获得更为准确的表达。

例句	领域	含义
新奥尔良与汉堡都是重要的港口城市，在海军力量投射和补给方面发挥着关键的作用。	军事、地理	两个城市
肯德基对新奥尔良烤鸡腿汉堡进行了改良。	食品	产品

代词的指代关系。虽然每种语言都有自己较为完善的语法规则，但在理解句子的过程中，仅依靠句法关系是不足以做出准确判断的，多数时候需要结合语义信息进行联合判断。这种情况在英文中出现频次较多，以下是案例。

英文例句	解释说明
Mary went to the bookstore to buy a new book. When she got home, she realized it was not the one she wanted.	从语法上来看，根据位置关系，可以推测 "it" 应该指代的是 "a new book"。然而，从语义上来看，我们知道购买的行为通常发生在书店而不是书本上，因此 "it" 更可能指代的是 "the bookstore"，而不是 "a new book"。

断句、重音、儿化音、多音字等。这些情况主要出现在中文文本中，以下例子给出了解释说明。这些不断变化的表达形式无一不给 NLP 任务带来了困难和挑战，但也正因如此，使得 NLP 在人工智能领域一直闪耀着璀璨的光辉。

中文例句	解释说明
问：豆腐多少钱？ 答：一块两块。	出现歧义：一块豆腐两块钱 or 一块钱两块豆腐
人要是行，干一行行一行，一行行行行行	多音字：行(háng)：行业 ； 行(xíng)：干得好 断句：一行(háng) / 行(xíng) / 行行(háng) / 行(xíng)
苯环中的碳碳键键能能否否定定律一	断句：碳碳键 / 键能 / 能否 / 否定/ 定律一
治理解放大道路面积水	正确断句：治理 / 解放 / 大道 / 路面 / 积水 错误断句：治 / 理解 / 放大 / 道路 / 面积 / 水