在整个片段中,“意思”一词在不同的语境里共有6个不同的含义。如果实现这个词义的自动理解,恐怕不是目前的自然语言处理系统所能够胜任的。当然,这个片段可能是人为编造出来的,实际运用中一般不会出现如此复杂的用词方法。我们使用这个例子的意思也绝不是说一个自然语言处理系统必须具备如此复杂的歧义消解能力才算得上是真正实用的系统,而只是想说明,歧义是自然语言中普遍存在的语言现象,它们广泛地存在于词法、句法、语义、语用和语音等每一个层面。任何一个自然语言处理系统,都无法回避歧义的消解问题。

另一方面,对于一个特定系统来说,总是有可能遇到未知词汇、未知结构等各种意想不到的情况,而且每一种语言又都随着社会的发展而动态变化着,新的词汇(尤其是一些新的人名、地名、组织机构名和专用词汇)、新的词义、新的词汇用法(新词类),甚至新的句子结构都在不断出现,尤其在口语对话或计算机网络对话(通过MSN、QQ、GTalk、Skype等形式)、微博、博客等中,稀奇古怪的词语和话语结构更是司空见惯。因此,一个实用的自然语言处理系统必须具有较好的未知语言现象的处理能力,而且有足够的对各种可能输入形式的容错能力,即我们通常所说的系统的鲁棒性(robustness)问题。当然,对于机器翻译、信息检索、文本分类等特定的自然语言处理任务来说,还存在若干与任务相关的其他问题,诸如如何处理不同语言的差异、如何提取文本特征等。

总而言之,目前的自然语言处理研究面临着若干问题的困扰,既有数学模型不够奏效、有些算法的复杂度过高、鲁棒性太差等理论问题,也有数据资源匮乏、覆盖率低、知识表示困难等知识资源方面的问题,当然,还有实现技术和系统集成方法不够先进等方面的问题。正是这些问题和困难,才使得自然语言处理研究更加充满挑战性,更需要我们去创新和探索。

1.3 自然语言处理的基本方法及其发展

1.3.1 自然语言处理的基本方法

一般认为,自然语言处理中存在着两种不同的研究方法,一种是理性主义(rationalist)方法,另一种是经验主义(empiricist)方法。

理性主义方法认为,人的很大一部分语言知识是与生俱来的,由遗传决定的。持这种观点的代表人物是美国语言学家乔姆斯基(Noam Chomsky),他的内在语言官能(innate language faculty)理论被广泛地接受。乔姆斯基认为,很难知道小孩在接收到极为有限的信息量的情况下,在那么小的年龄如何学会了如此之多复杂的语言理解的能力。因此,理性主义的方法试图通过假定人的语言能力是与生俱来的、固有的一种本能来回避这些困难的问题。

在具体的自然语言问题研究中,理性主义方法主张建立符号处理系统,由人工整理和编写初始的语言知识表示体系(通常为规则),构造相应的推理程序,系统根据规则和程序,将自然语言理解为符号结构——该结构的意义可以从结构中的符号的意义推导出来。按照这种思路,在自然语言处理系统中,一般首先由词法分析器按照人编写的词法规则对

输入句子的单词进行词法分析，然后，语法分析器根据人设计的语法规则对输入句子进行语法结构分析，最后再根据一套变换规则将语法结构映射到语义符号（如逻辑表达式、语义网络、中间语言等）。

而经验主义的研究方法也是从假定人脑所具有的一些认知能力开始的。因此，从这种意义上讲，两种方法并不是绝对对立的。但是，经验主义的方法认为人脑并不是从一开始就具有一些具体的处理原则和对具体语言成分的处理方法，而是假定孩子的大脑一开始具有处理联想（association）、模式识别（pattern recognition）和通用化（generalization）处理的能力，这些能力能够使孩子充分利用感官输入来掌握具体的自然语言结构。在系统实现方法上，经验主义方法主张通过建立特定的数学模型来学习复杂的、广泛的语言结构，然后利用统计学、模式识别和机器学习等方法来训练模型的参数，以扩大语言使用的规模。因此，经验主义的自然语言处理方法是建立在统计方法基础之上的，因此，我们又称其为统计自然语言处理（statistical natural language processing）方法。

在统计自然语言处理方法中，一般需要收集一些文本作为统计模型建立的基础，这些文本称为语料（corpus）。经过筛选、加工和标注等处理的大批量语料构成的数据库叫做语料库（corpus base）。由于统计方法通常以大规模语料库为基础，因此，又称为基于语料（corpus-based）的自然语言处理方法。

实际上，理性主义和经验主义试图刻画的是两种不同的东西。Chomsky 的生成语言学理论试图刻画的是人类思维（I-language）的模式或方法。对于这种方法而言，某种语言的真实文本数据（E-language）只是提供间接的证据，这种证据可以由以这种语言为母语的人来提供。而经验主义方法则直接关心如何刻画这些真实的语言本身（E-language）。Chomsky 把语言的能力（linguistic competence）和语言的表现（linguistic performance）区分开来了。他认为，语言的能力反映的是语言结构知识，这种知识是说话人头脑中固有的，而语言表现则受到外界环境诸多因素的影响，如记忆的限制、对环境噪声的抗干扰能力等。

1.3.2 自然语言处理的发展

理性主义和经验主义在基本出发点上的差异导致了在很多领域中都存在着两种不同的研究方法和系统实现策略，这些领域在不同的时期被不同的方法主宰着。

在20世纪20年代到60年代的近40年时间里，经验主义方法在语言学、心理学、人工智能等领域中处于主宰的地位，人们在研究语言运用的规律、言语习得、认知过程等问题时，都是从客观记录的语言、语音数据出发，进行统计、分析和归纳，并以此为依据建立相应的分析或处理系统。

大约从20世纪60年代中期到20世纪80年代中后期，语言学、心理学、人工智能和自然语言处理等领域的研究几乎完全被理性主义研究方法控制着，人们似乎更关心关于人类思维的科学，人们通过建立很多小的系统来模拟智能行为，这种研究方法一直到今天还仍然有人在使用。但是，这种做法常常受到批评，因为这种做法只能处理一些小的问题，而不能对研究方法的有效性给出一个总的客观的评估，因此，这些小系统有时也被轻蔑地称为玩具[Manning and Schütze,1999]。