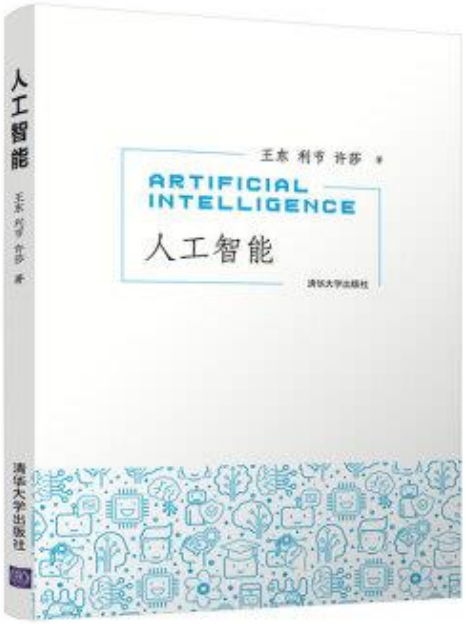
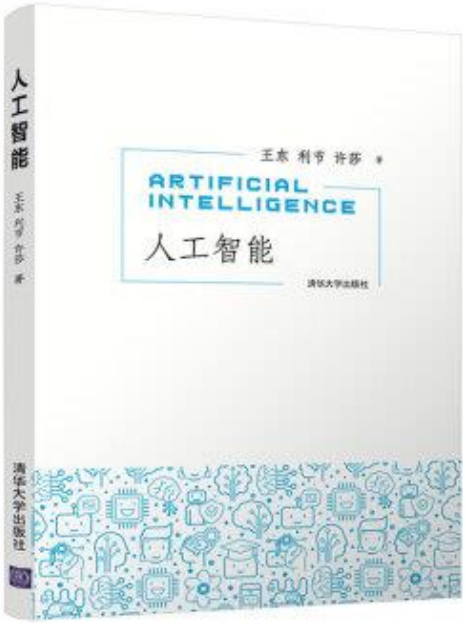
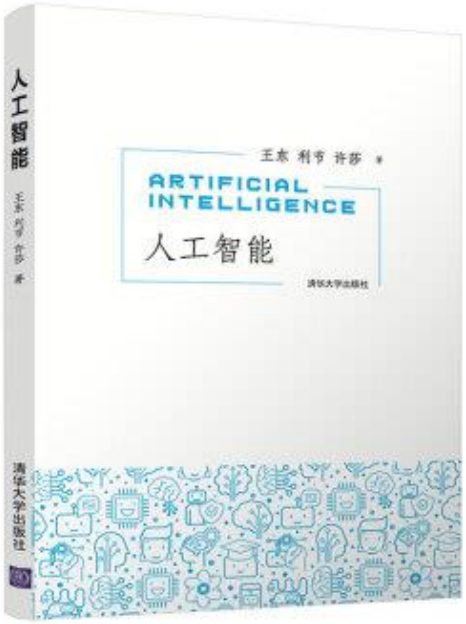


理解你的语言

利节

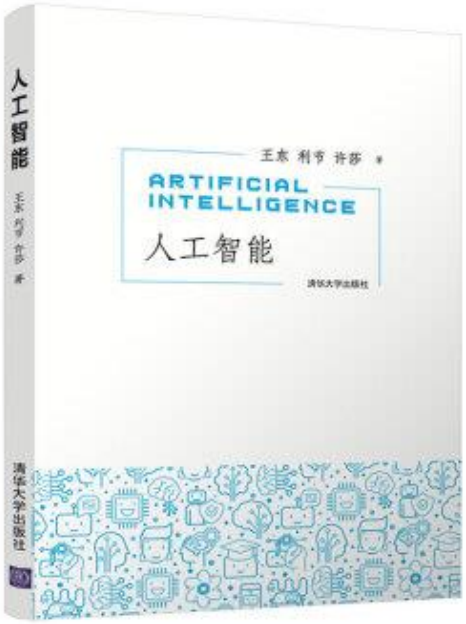


语言是人类特有的能力，集中体现了人的智能性。据统计，地球上有将近7000 种语言，其中有2000 多种语言有书面文字。不论是否有文字，每种语言都有其独特的发音方式、组词规则、句法结构等，表现出非常复杂的各异性。人工智能的研究者一直把理解和掌握人类语言作为实现机器智能的重要目标。然而，人类语言是如此复杂，即使是人都很难掌握（大家可以回忆自己学习外语时的痛苦经历），让机器对其完全理解是不太可能的。然而，在一些特定领域，对语言进行部分、浅层的理解是可能的。



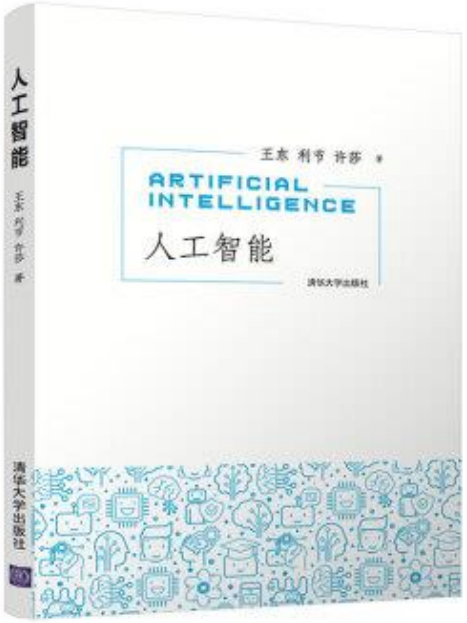
目 录

- 人类语言的复杂性
- 传统语言理解方式
- 基于深度学习的语言理解方法
- 机器翻译
- 语言理解的其他应用



目 录

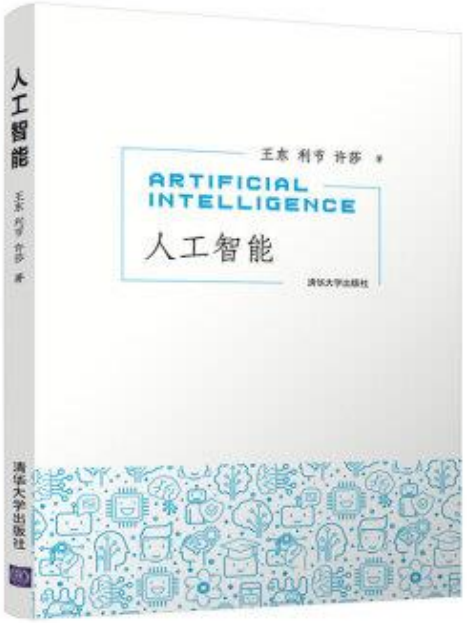
- 人类语言的复杂性
- 传统语言理解方式
- 基于深度学习的语言理解方法
- 机器翻译
- 语言理解的其他应用



人类语言的复杂性

1)结构复杂性

人们在用语言进行意思表达的时候，首先会在脑海中形成需要表达的意图（Intention），依据这些意图选择合理的表达方式（如因果方式，总分方式等）；在表达中的每一句话都有一个核心思想，基于此选择合适的词，并依据语法规则将这些词有序地组织起来，即形成了一段包含书写人意图的段落。



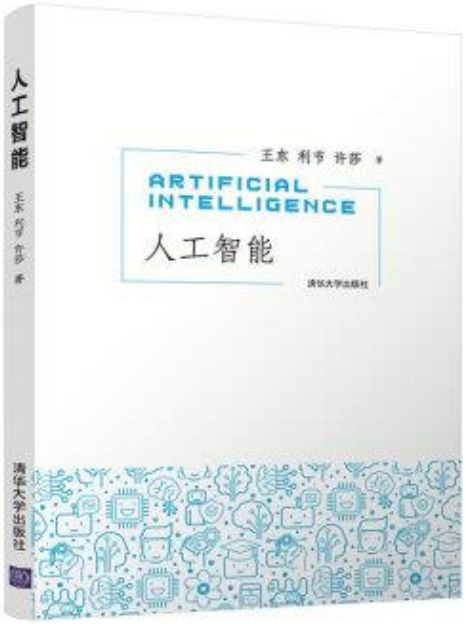
人类语言的复杂性

1)结构复杂性

例如下面这段话：

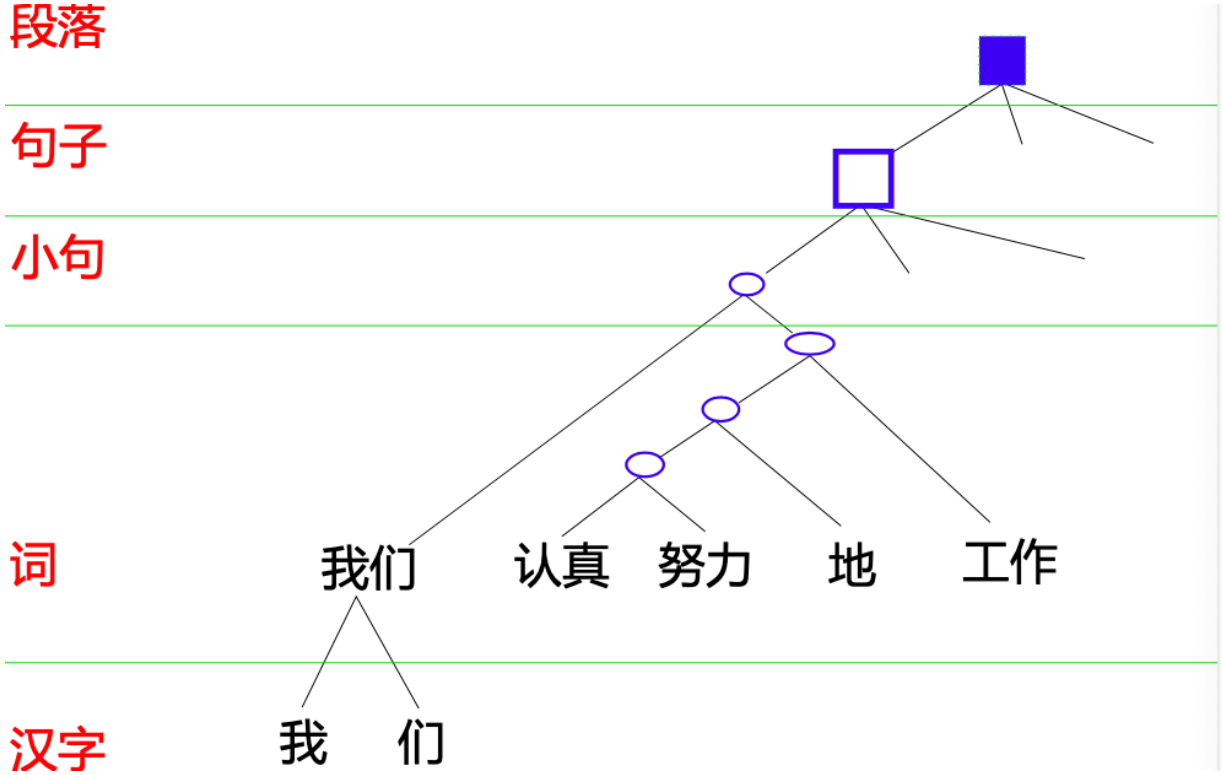
我们认真努力地工作，就会做出更大贡献，让我们的国家更强大。否则，别的国家就会看不起我们。

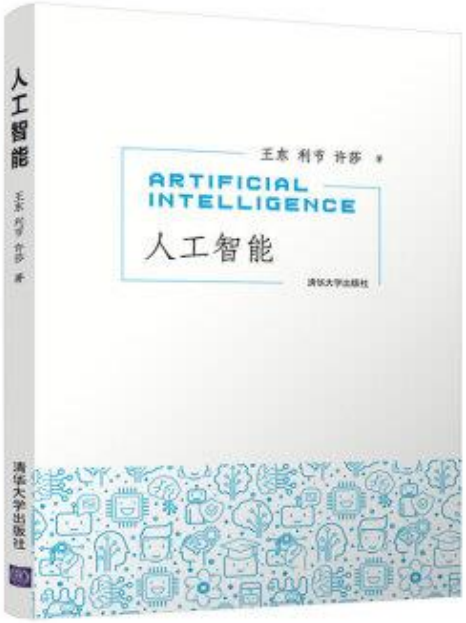
上面这段话没有任何生僻字，但包含丰富的说理过程。我们试着对该段话的表达方式进行分析



人类语言的复杂性

1)结构复杂性





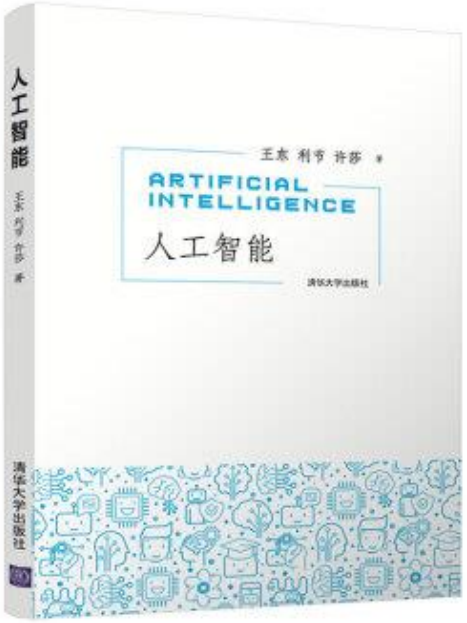
人类语言的复杂性

2)语义复杂性

语义复杂性一方面来源于词本身的歧义性，也来源于词与词互相组合时的结构歧义性。
例如下面这句话：

我想起来了

既可以理解成“我/想起来/了”，也可以理解成“我想/起来了”。这里的歧义性既源于“想”和“起来”这个两词本身的歧义，也源于“起来”和“想”这两个词的结合方式。



人类语言的复杂性

2)语义复杂性

另一个例子：

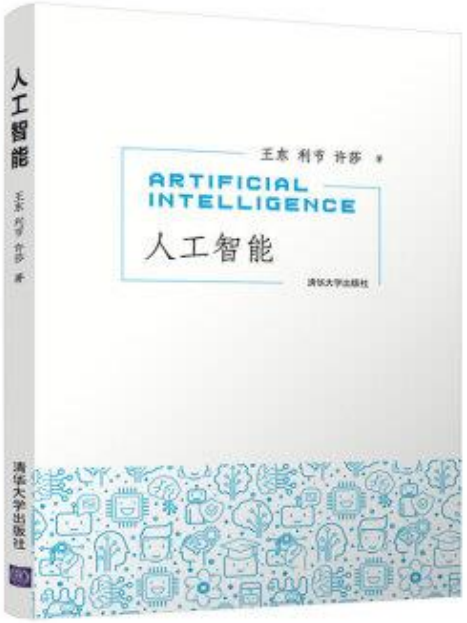
他在那儿看东西

既可以理解成“看守东西”，也可以理解成“用眼睛看东西”。这里的歧义性在口语中是不会发生的，但写成文字就会出现。

再举一个例子：

我要榨果汁

可以理解成“我想要自己榨杯果汁”，也可以理解成“我想点一杯榨果汁”。这里的歧义主要来源于“榨果汁”三个字在组词时的多义性。

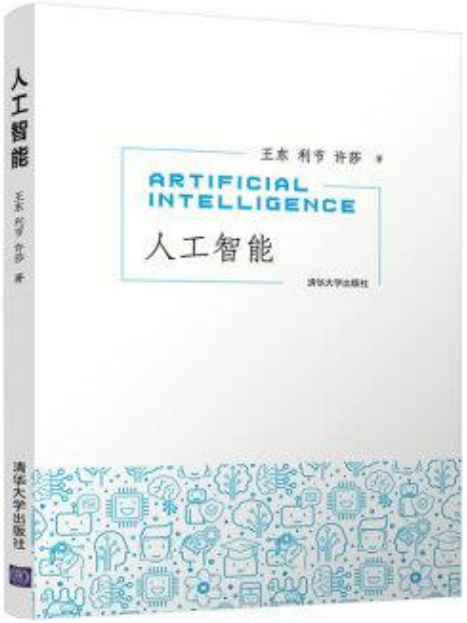


人类语言的复杂性

3)知识复杂性

语言之所以复杂的另一个原因是语言中不仅包含语法结构和语义内容，还包含丰富的知识。例如下面这句话：

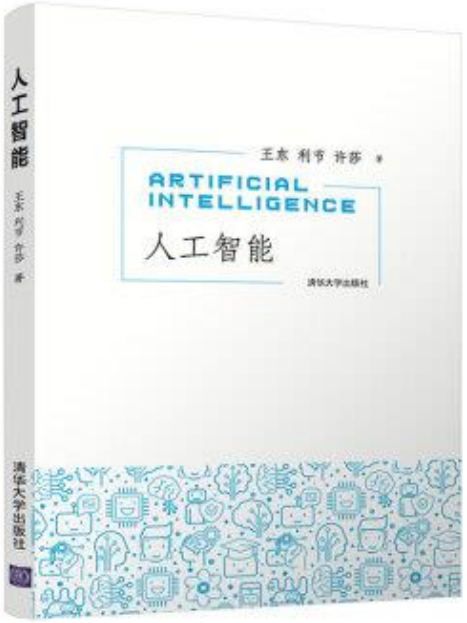
煤中含碳元素，不充分燃烧时会产生一氧化碳。一氧化碳进入人体和血红细胞结合，使血红细胞失去携氧功能，产生一氧化碳中毒。因此，我们在烧煤时要尽量保持通风，特别是在低压天气，要特别注意。



人类语言的复杂性

4)时空复杂性

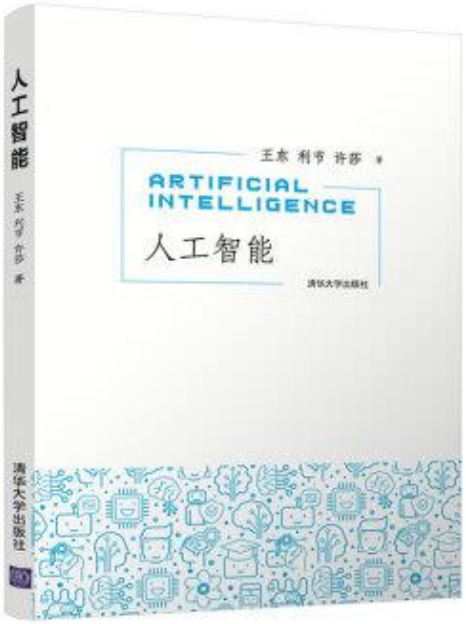
全球有近7000 种语言，每种语言都有其独特之处。这些语言被分成不同语系和语族，不同种类的语言差异巨大。即使同一种语言，语言习惯也会随地域的不同而不同。典型的如大陆中文和台湾中文，英国英语和美国英语等。另外，同一种语言在不同历史时期在遣词造句方面也有很大差异，典型的如古代汉语和现代汉语，中世纪英语和现代英语等。



人类语言的复杂性

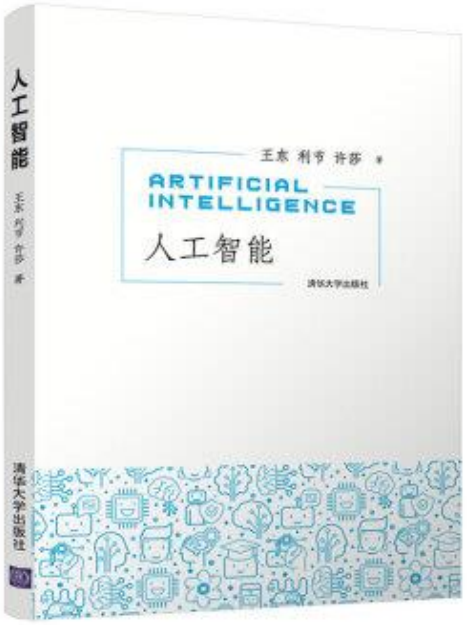
5)应用复杂性

标准的书面语相对比较规范，但在一些实际场景下（如微信、论坛、微博等），人们所使用的语言很多时候是不规范的，拼写错误、句法错误等很常见。即便是正规出版物，出错也不可避免。如2006 年有篇新闻标题为《消防安全隐患构建和谐粮库》，明显有语义不通的问题。这些不规范用语使得语言理解任务更为复杂。特别重要的是，传统语言理解方法是基于句法分析的，其前提是句子遵循合理的句法，如果句法不通，再强大的分析工具都无能为力。



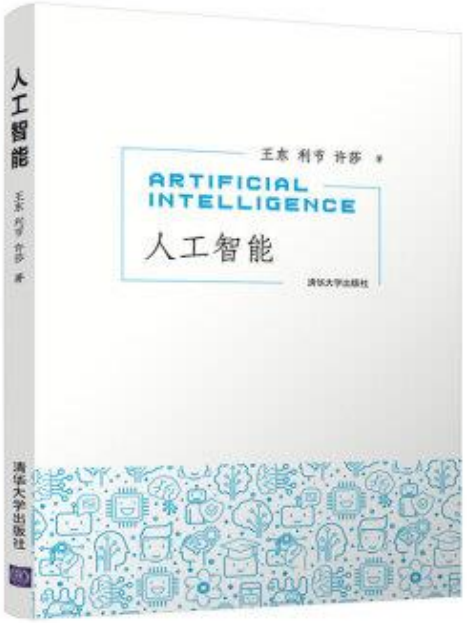
目 录

- 人类语言的复杂性
- 传统语言理解方式
- 基于深度学习的语言理解方法
- 机器翻译
- 语言理解的其他应用



传统语言理解方式

传统语言理解方法以句子分析为基本出发点，通过分析句子中的词法、句法、语义，实现对一句话的细致拆解。我们将从词法分析、句法分析、语义分析三个层次介绍传统语言理解方法。

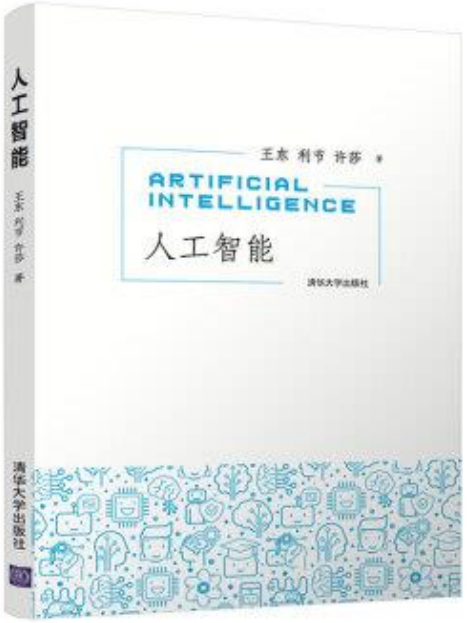


传统语言理解方式

1)词法分析

词是最小的表达单元。对汉语来说，词可能只有一个字，如“走”、“跑”、“花”等，也可包含多个字，如“打球”，“上树”等。每个词既是一个语义单元，也是一个语法单元。

所谓词法分析，是指从输入序列中确定词序列，并标记每个词的词性。因为汉语没有明确的词边界，因此词法分析首先需要将连续的汉字序列切分成独立的词，这一过程称为“分词”。分词是汉语独特的词法分析任务。



传统语言理解方式

1)词法分析

下面这句话：

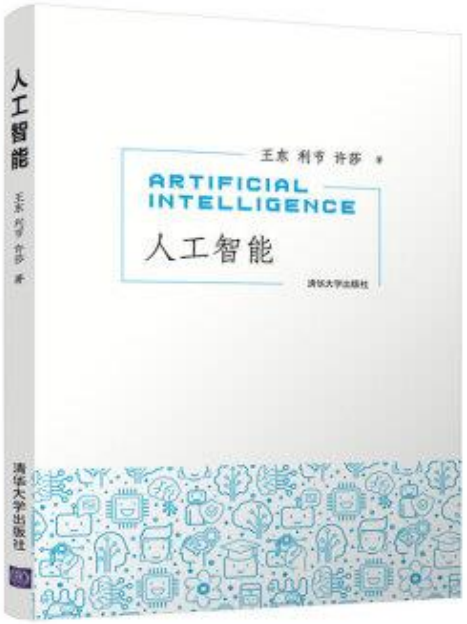
中华民族是伟大的民族

需要切分成

中华/民族/是/伟大/的/民族

另一方面，所有的词都是由字串组成的，这意味着分词具有一定的灵活性，如上面这句话也可拆分成如下形式：

中华民族/是/伟大的/民族



传统语言理解方式

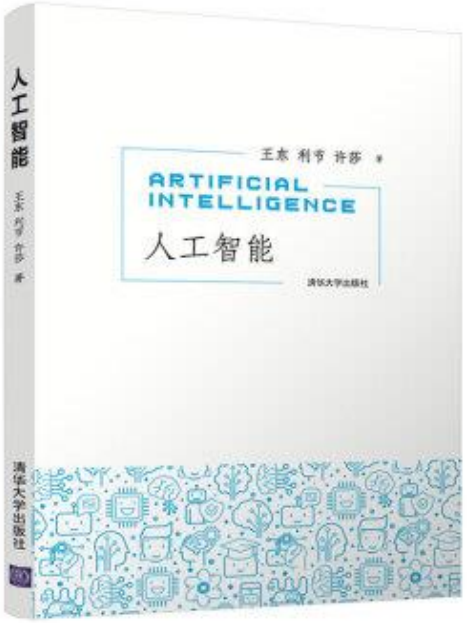
1)词法分析

汉语分词有多种方法，最常见的是最长匹配法，即尽可能用词表中最长的词对句子进行切分。值得说明的是，不同切分方法有可能带来语义上的差别。

例如对“发现大道有活动”这句话进行下面两种切分，会产生完全不同的意义：

发现/大道/有/活动

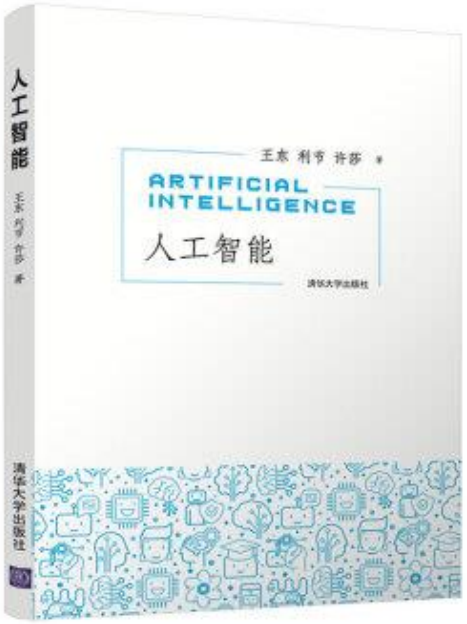
发现大道/有/活动



传统语言理解方式

1)词法分析

汉语词法分析的另一个重要任务是对词性进行标注，如将“吃/黄瓜/好”标注成“动词/名词/形容词”。词性标注算法有很多，较早的方法依赖规则，如动词后一般接名词，名词前多出现形容词等。当前的词性标注方法多基于统计模型，特别是HMM模型、最大熵模型等，这些模型统计不同词性的单词相互连接的概率，基于该模型，在标注时选择最大概率的词性序列。



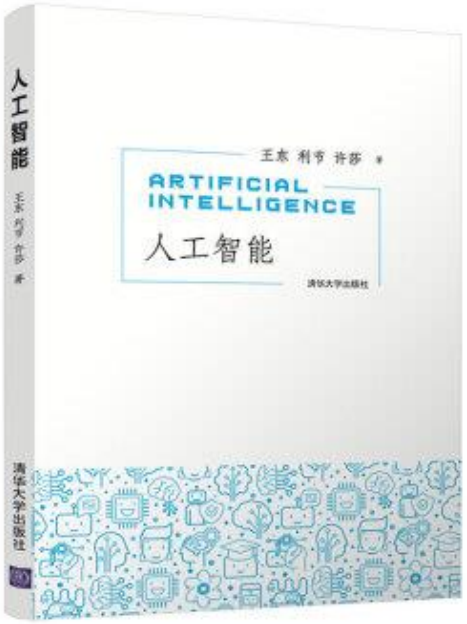
传统语言理解方式

2)句法分析

句法分析是在词法分析的基础上，对一句话中词与词的组合方式进行解析。常见的句法分析有两种：

（1）成分结构分析，用以分析句子的层次性组织结构；

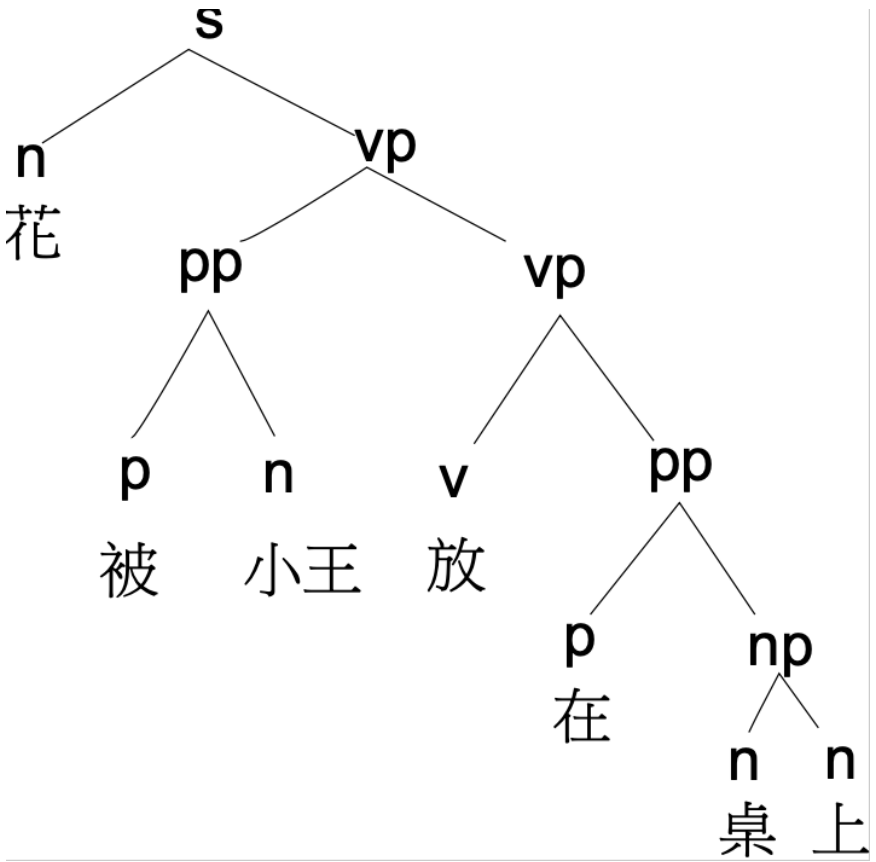
（2）依存分析，用以分析词与词之间的互相依赖性。

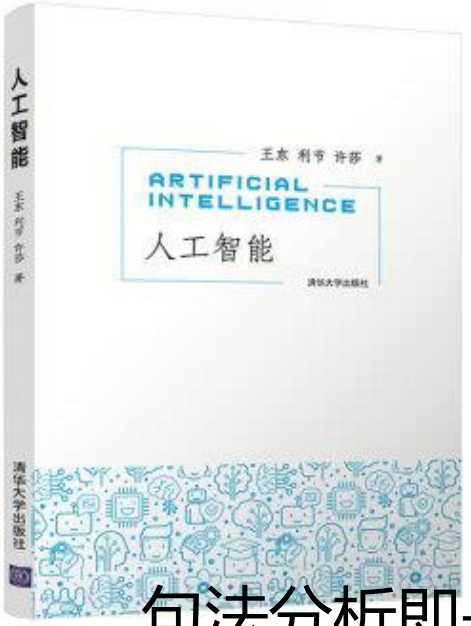


传统语言理解方式

2)句法分析

词与词之间通过层次性结构组合成句子，可以将句子表示为一棵句法树。在这棵句法树中，每个叶子节点对应一个词，词与词之间组成短语，构成低层的中间节点，这些中间节点再互相组合，形成更高层次的中间节点，对应更大规模的短语。这一组合方法迭代进行，直到形成一棵完整的句法树。





传统语言理解方式

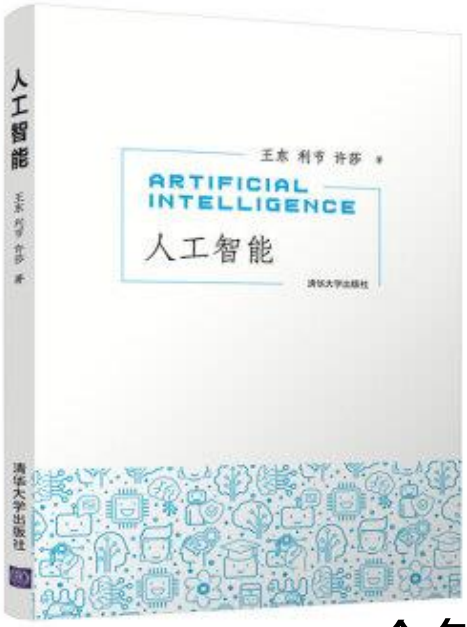
2)句法分析

句法分析即是将一句话分解成句法树的过程。传统句法分析基于规则。这一方法假设我们所用的自然语言是一个完美的语法系统，这套语法系统可表示成一套生成规则，语言中的所有句子可由这些规则生成。一种常用的语法系统具有如下形式：

A → BC;

A → a

其中A,B,C 表示任一个非终结符，代表名词短语、动词短语等，而a 为任意一个终结符，代表一个单词。基于这一语法即可生成一个完整的句子，而每句话都可表示为一个句法树。这种形式的生成规则称为上下文无关文法（Context Free Grammar，CFG）。



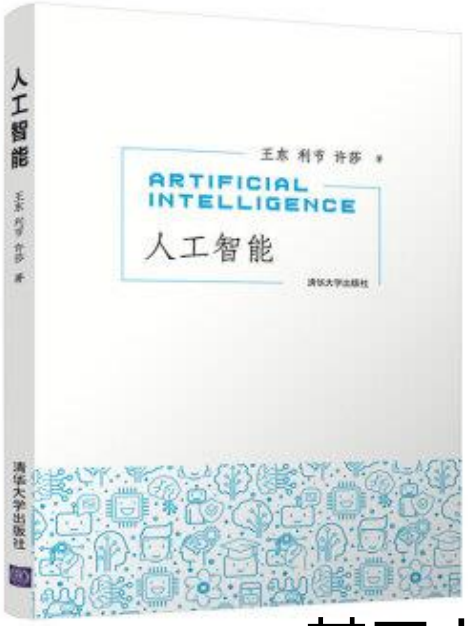
传统语言理解方式

2)句法分析

一个简单CFG 的例子如下（我们称为水果CFG）：

规则1: S → N VP;
规则2: VP → V N;
规则3: V → 吃| 拿
规则4: N → 猴子| 苹果| 香蕉

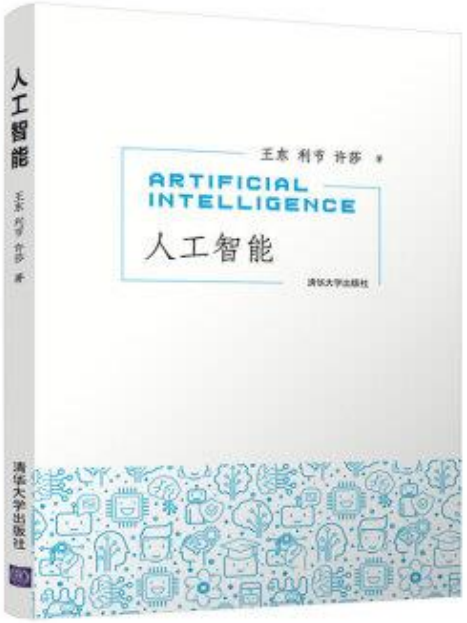
其中S 代表句子，N 为名词，V 为动词，VP 为动词词组。这一水果CFG可生成“猴子吃苹果”、“苹果拿香蕉” 等简单句子。



传统语言理解方式

2)句法分析

基于上下文无关文法，可以对句子进行句法分析，构造句法树。分析方法是对句子从左到右扫描，选择合适的生成规则对词和短语进行合并，得到更高层的短语，这一过程称为逆向推理。如果我们可以找到一系列生成规则及其应用顺序，使得目标句子得以生成，即可获得该句对应的句法树，实现了对这句话的句法分析。

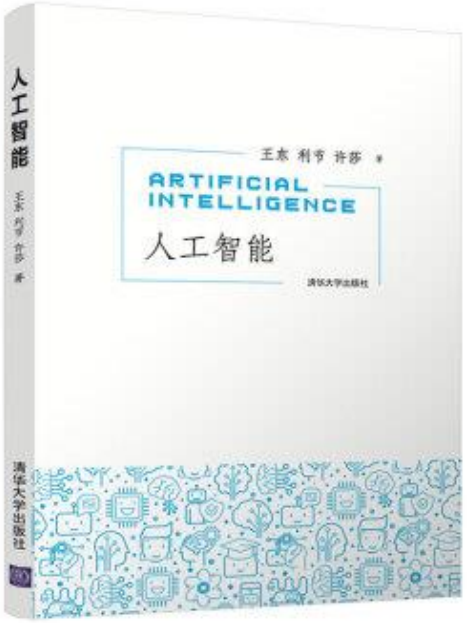


传统语言理解方式

2)句法分析

在上面水果CFG 的例子中，如果给定一个句子“猴子吃香蕉”，首先经过分词后得到词序列“猴子/吃/香蕉”。对这一序列从左到右扫描，可得到如下逆向推理过程：

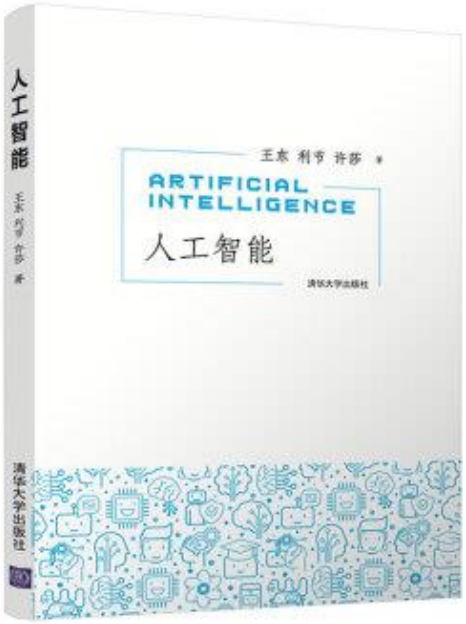
- 步骤1: 猴子→ N （规则4）；
- 步骤1: 吃→ V （规则3）；
- 步骤2: 香蕉→ N （规则4);
- 步骤3: V N → VP （规则2);
- 步骤3: N VP → S （规则1）



传统语言理解方式

3)语义分析

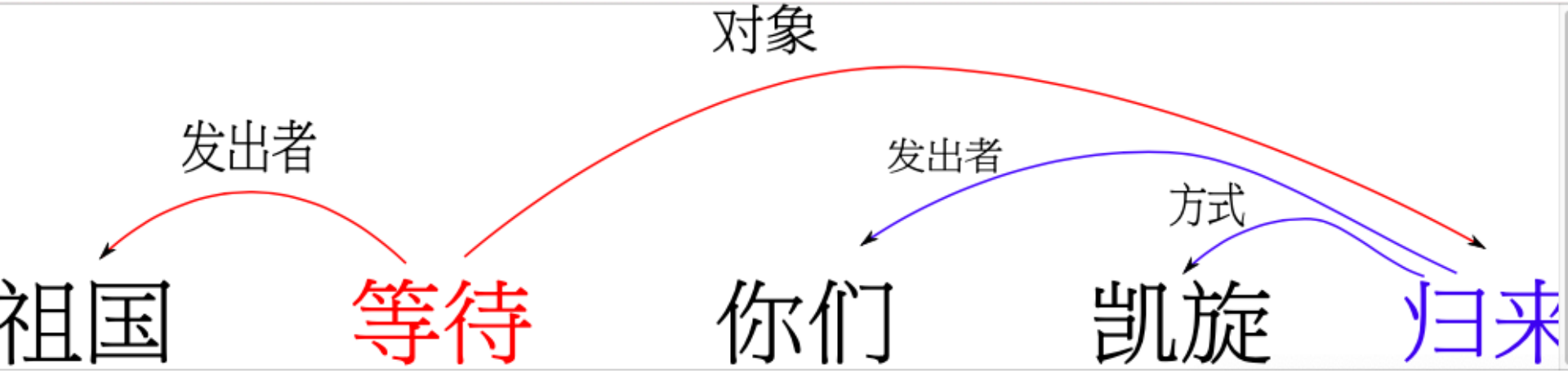
一种简单的语义分析方法是找到一句话中的中心谓词，确定该谓词的相关成分，如施事、受事、时间和地点等。这一过程称为语义角色标注，也称浅层语义分析。更完整的语义分析是对句子的各个成分进行细致解析，形成某种形式化表示。一个好的形式化表示方法应具有如下性质：第一，它表达出的语义应该是格式化的、简洁的。第二，它表达出来的意义应该是明确的，没有歧义的。第三，它的表达必须足够灵活，足以表示较复杂的语义。

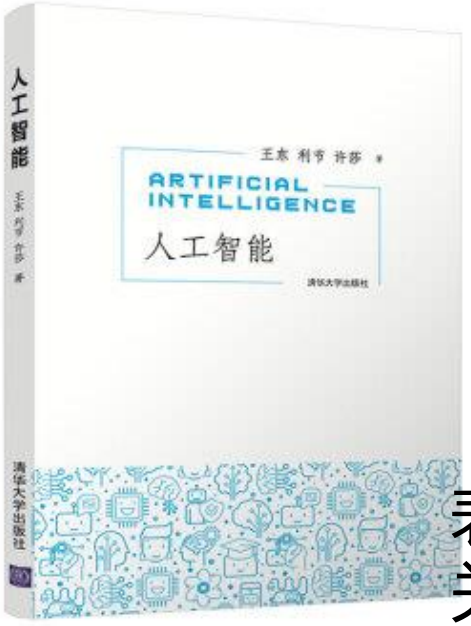


传统语言理解方式

3)语义分析

语义依存树是一种常用的形式化表示方法，该树表达了句子中各个组成成分所代表的语义角色及其相互关系。





传统语言理解方式

3)语义分析

拥有（我, 自行车）

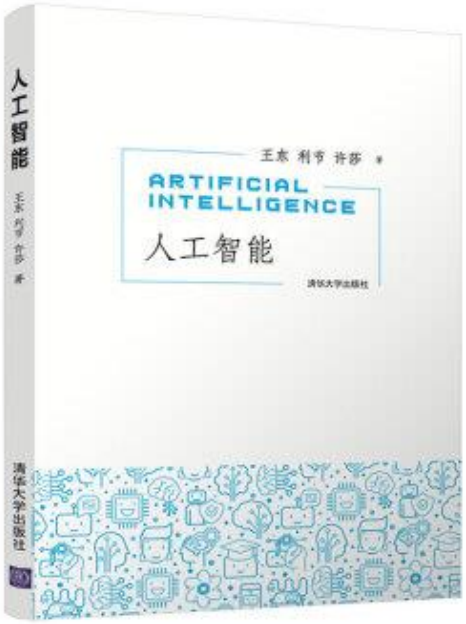
表达了“我拥有自行车” 这一事实，其中“拥有” 是关系，“我” 和“自行车”分别是这一关系相关的两个概念。如果我们能构造这一个三元组集合（人为的或自动抽取的），即可以建立一个知识库，如：

拥有（我, 自行车）
拥有（小明, 汽车）
喜欢（张亮, 小云）

基于这一知识库，即可回答“张亮喜欢谁” 这样的问题。具体步骤如下：首先，对问题进行语义分析，将分析结果表示为三元组，即：

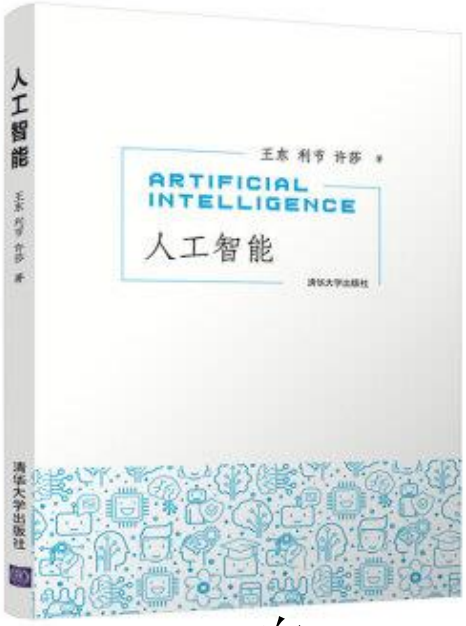
喜欢（张亮，?）

对这一表达在知识库中搜索，即可发现张亮喜欢的是“小云”。



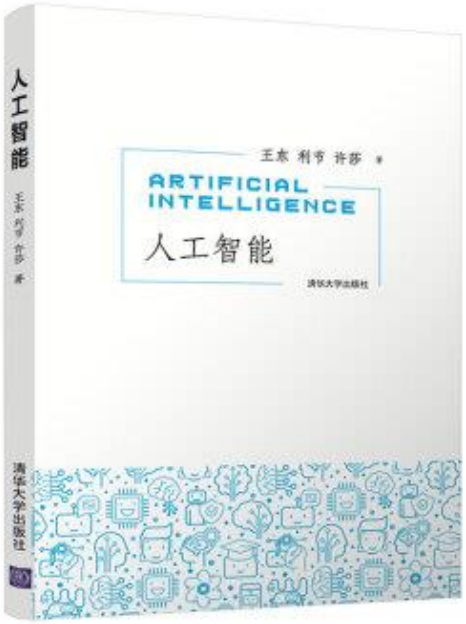
目 录

- 人类语言的复杂性
- 传统语言理解方式
- 基于深度学习的语言理解方法
- 机器翻译
- 语言理解的其他应用



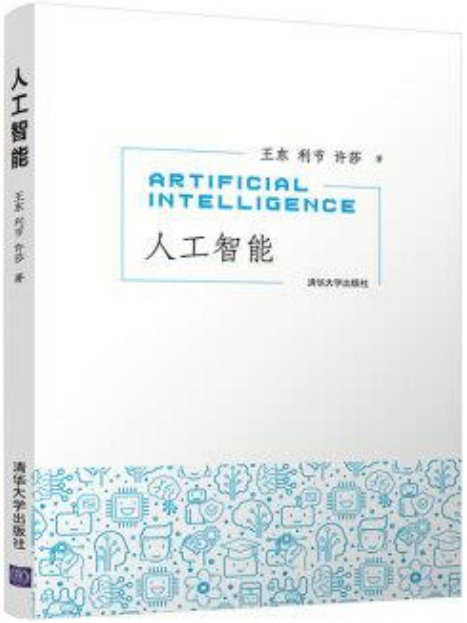
基于深度学习的语言理解方法

自2013 以来，基于深度学习的语言理解方法受到越来越多关注。和传统方法不同，这一方法并不依赖对句子成分的解析，而是将句子映射到一个语义空间里。在这个空间里，语义相近的句子距离较小，语义相差较大的句子距离较大。基于这一语义空间，可以实现问答、翻译等自然语言处理任务。值得说明的是，基于这种语言理解方法，机器对句子并没有一个非常明确的“理解”过程，但确实可以完成很多自然语言理解任务。因此，这一方法事实上是基于语言理解的反馈思路：只要可以顺利完成目标任务，则认为机器已经具有了理解能力。



目 录

- 人类语言的复杂性
- 传统语言理解方式
- 基于深度学习的语言理解方法
- 机器翻译
- 语言理解的其他应用

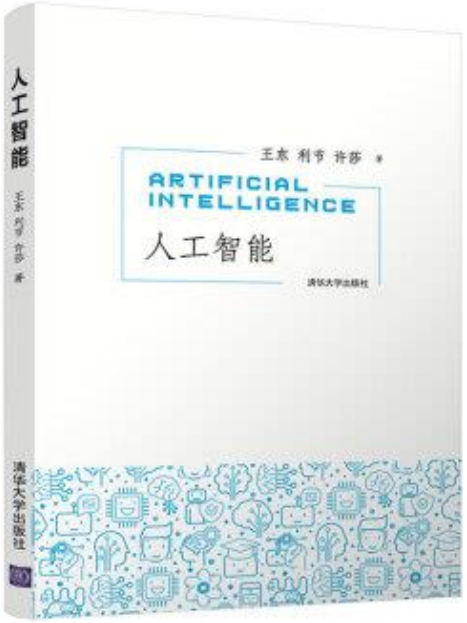


机器翻译

1)机器翻译的历史

Warren Weaver 在1947 年写给Norbert Wiener 的信中就谈到了机器翻译的设想。

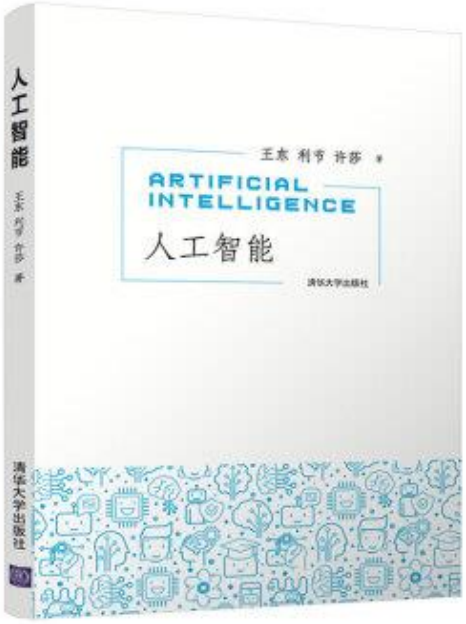
70年代后期，由加拿大蒙特利尔大学与加拿大联邦政府翻译局联合开发的TAUM-METEO 系统，可用来翻译气象预报。



机器翻译

1)机器翻译的历史

一直到80 年代末期，机器翻译的主要方法都是基于规则的。这些系统一般需要一本字典和一些语言学规则。在翻译时，首先通过查字典把每个单词在目标语言里对应的词找到，再依目标语言的语言学规则对这些词进行调整（形态、顺序等），最后组合成句子完成翻译。这一方法称为直接翻译法。

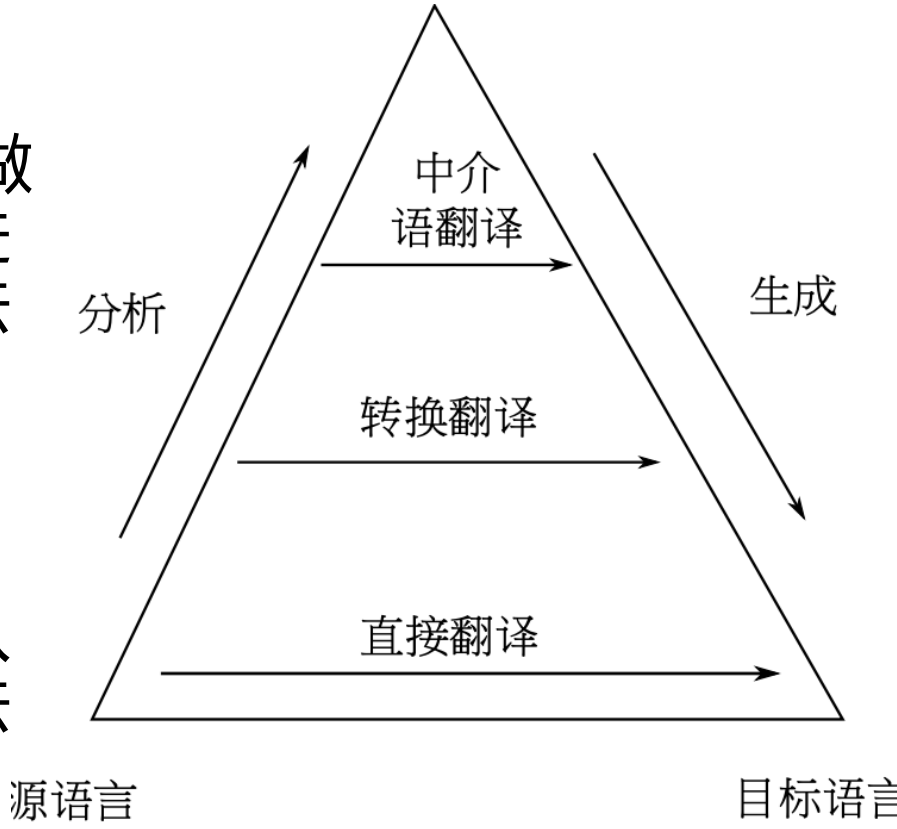


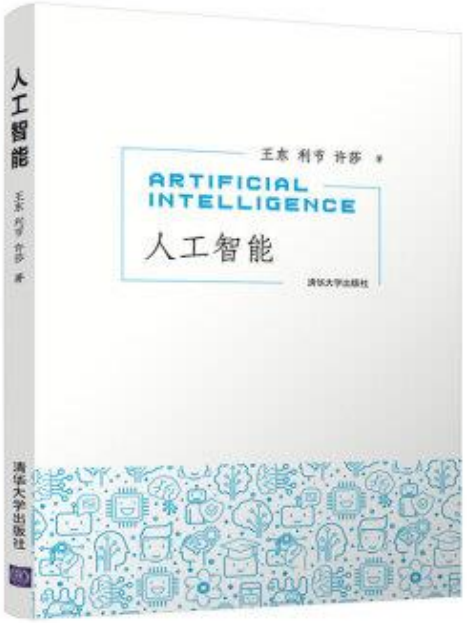
机器翻译

1)机器翻译的历史

另一些研究者提出应该对原句做语法分析，再对得到的语法成分进行翻译。这一方法称为转换翻译法。

还有学者尝试将源语言转化为一种表达抽象意义的中介语，再从中介语转换到目标语言。这一方法称为中介语翻译法。





机器翻译

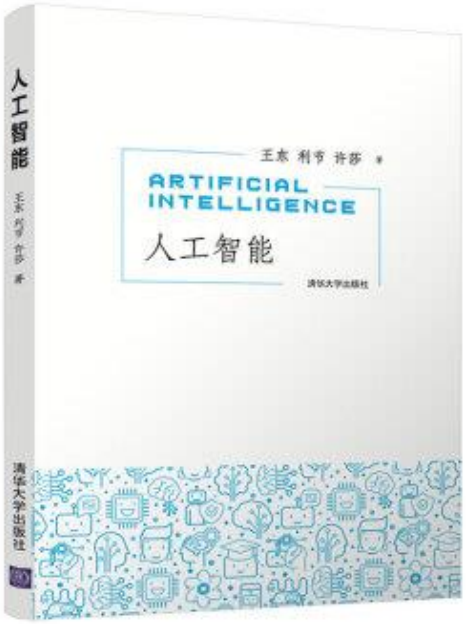
1)机器翻译的历史

突破发生在1993 年，IBM 的Brown 和Della Pietra 等人提出了基于词对齐的翻译模型，标志着现代统计机器翻译（SMT）方法的诞生。

2014 年，Google 的研究者提出基于神经网络的翻译模型，标志着神经机器翻译（NMT）的起点。

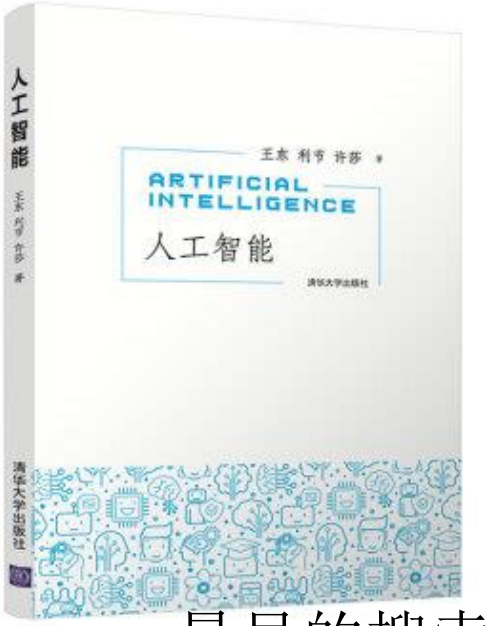
2016 年9 月，谷歌宣布其NMT 系统取得了逼近人类的性能。

2018 年3 月，微软宣布其NMT 系统在中英翻译上超过人类。



目 录

- 人类语言的复杂性
- 传统语言理解方式
- 基于深度学习的语言理解方法
- 机器翻译
- 语言理解的其他应用

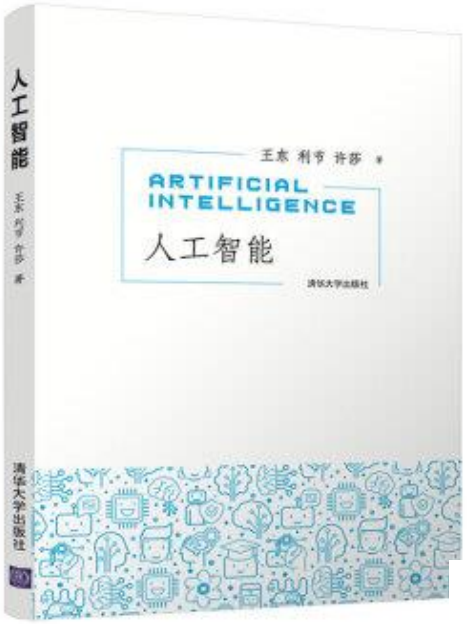


语言理解的其他应用

1)搜索引擎

最早的搜索引擎可能是1990 年的Archie 系统，这一系统是针对FTP服务资源的搜索器。1993 年，第一个面向网页的搜索引擎World Wide Web Wanderer 出现，同年，AliWeb 诞生。这些早期搜索引擎的功能有限，如AliWeb 只能对网页标题进行索引。1994 年创立的InfoSeek 推出搜索服务，后推出网景浏览器。同年，杨致远和David Filo 创立Yahoo!，该公司迅速成为搜索领域的主力军。

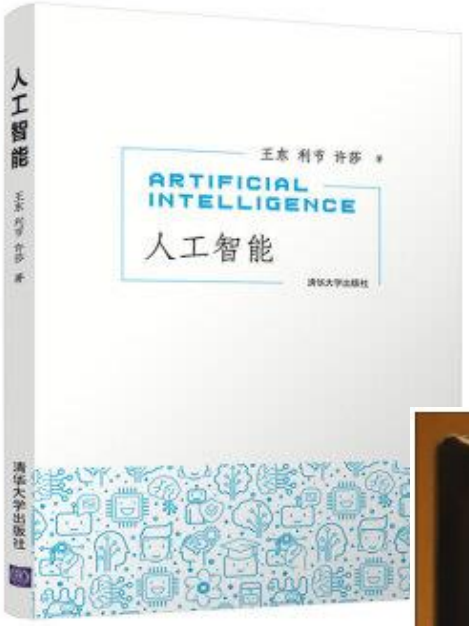
近年来，个性化搜索受到更多关注。这种搜索方法通过分析用户的搜索历史来发现用户的搜索倾向，依此对搜索结果进行调整，可得到更好的用户体验。



语言理解的其他应用

2)推荐系统

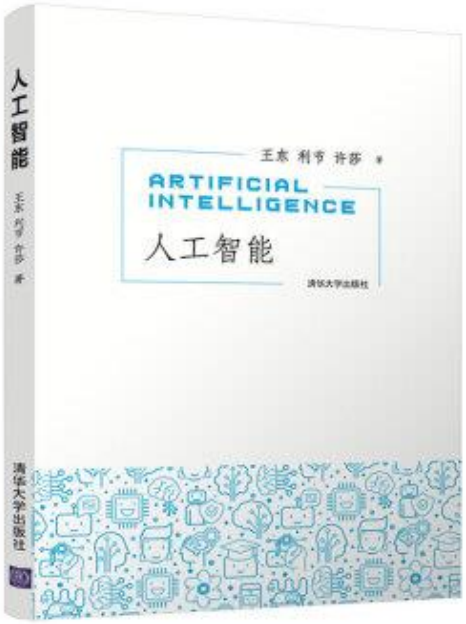
推荐是指我们在搜索某一内容时，系统自动为我们推荐相关的其他内容。



语言理解的其他应用

3)会说话的机器人

https://www.youtube.com/watch?v=iqHAL_8Ug2Y



The end !