



目录

CONTENTS

10.1

计算机视觉概述

10.2

数字图像的类型及表示

10.3

常用计算机视觉模型和关键技术

10.4

应用实例：人脸识别技术





中國石油大學(华东)
CHINA UNIVERSITY OF PETROLEUM

计算机科学与技术学院
Big Data & Cyber Security

第十一章 自然语言处理

人工智能课程组

智能科学系





目录

CONTENTS

11.1 自然语言处理概述

11.2 词法分析

11.3 句法分析

11.4 语义分析

11.5 词库

11.6 机器翻译

11.7 自然语言人机交互

11.8 智能问答





11.1 自然语言处理概述

- 人工智能分为三个层次：
 - 运算智能：记忆和计算
 - 感知智能：感知环境的能力，如听觉、视觉、触觉等
 - 认知智能：语言理解，知识和推理
 - 创造智能：对未见过、未发生过的事物和问题，运用经验，通过想象力设计、实验、验证并予以实现的智力过程。
- 自然语言处理是从自然语言到机器内部的一个映射
 - ✓ 自然语言处理是认知智能的核心
 - ✓ 引导知识图谱进步
 - ✓ 提高用户理解能力
 - ✓ 推动推理能力

11.1 自然语言处理概述

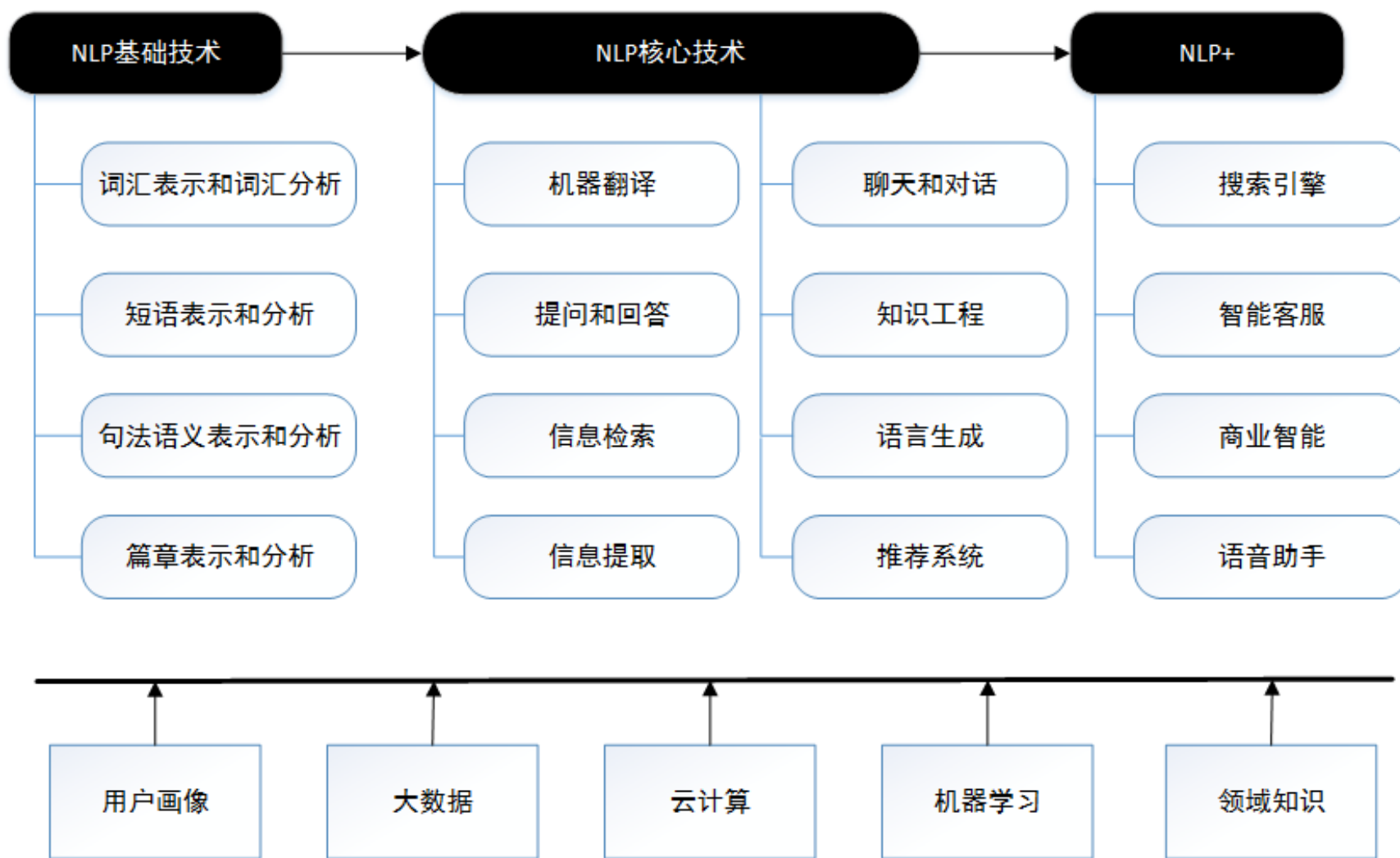


图11.1 自然语言处理框架图

11.1 自然语言处理概述

- 自然语言处理的基本任务:
- 分词模块: 将汉字序列切分为单词序列
- 词性标注: 为每个单词标注词性
- 依存句法分析: 预测句子中的单词之间的依存关系, 用树结构表示句法结构
- 命名实体识别: 从文本中识别具有特殊意义的实体

句子输入:

我爱自然语言处理

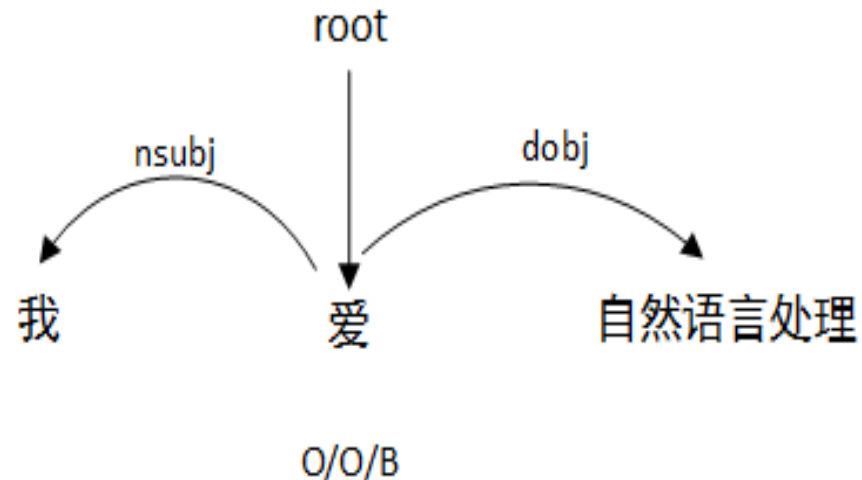
句子输出:

我/爱/自然语言处理

词性标注输出:

PN//VV/NR

依存句法分析输出:



命名实体识别输出:

图11.2 自然语言处理示例

11.1 自然语言处理概述

- 自然语言处理的历史
- 第一阶段（60-80年代）：基于规则建立词汇、句法语义分析、问答、聊天和机器翻译系统，
 - 规则可以利用人类知识，不依赖数据；覆盖面不足，规则关联和可扩展一直没有解决
- 第二阶段（90年代开始）：基于统计的机器学习
 - 利用带标注的数据，基于人工定义的特征建立机器学习系统
- 第三阶段（2008年之后）：深度学习
 - 将深度学习用于特征计算

11.1 自然语言处理概述

- 自然语言处理中的深度学习技术
 - 神经网络端对端训练
 - 使用标注数据（如双语对照语料）训练神经网络模型
 - 词嵌入（word embedding）
 - 对大规模语料训练获得多维语义空间上的表达，定义词汇之间、短语之间、句子之间、篇章之间的语义距离
 - 基于神经网络的语言模型
 - 预测下一个词或者下一个句子
 - 循环神经网络
 - 对不定长的句子编码
 - 编码-解码技术
 - 一个句子到另一句子的变化，翻译、问答、转述
 - 强化学习
 - 通过用户或者环境的反馈调整神经网络的参数

11.1 自然语言处理概述

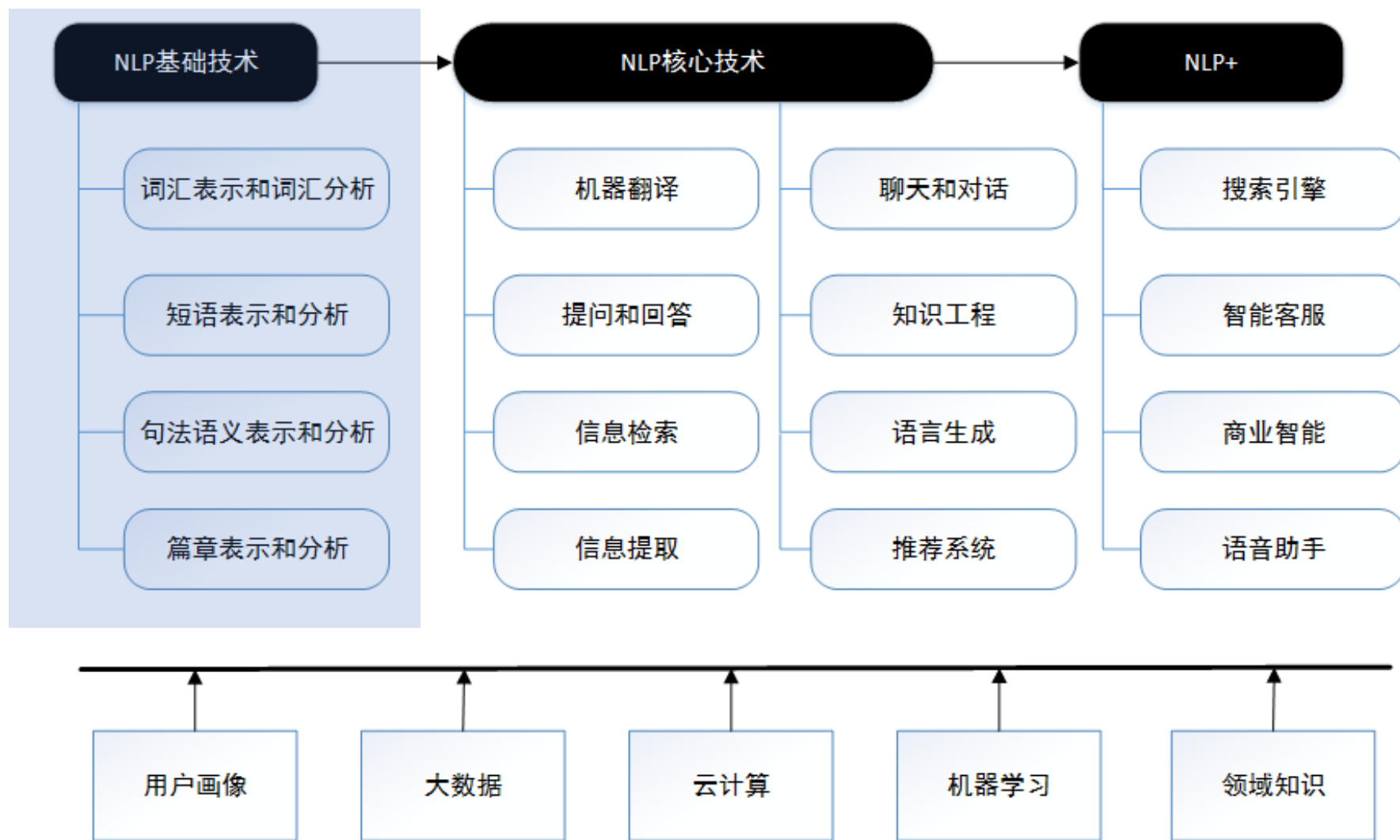


图11.1 自然语言处理框架图

11.2 词法分析

- 定义：从句子中**切分**出单词，找出词汇的各个**词素**，并确定其词义。

例：unchangeable: un-change-able

■ 英语词法分析

特点：切分单词容易，找出词素复杂。

例：importable分为 import-able或 im-port-able

词法分析算法举例：

repeat

look for word in dictionary

if not found

then modify the word

Until word is found or no further modification possible

11.2 词法分析

例如：对于单词catches、ladies可以做如下的分析。

catches	ladies,	词典中查不到
catche	ladie	修改1: 去掉s
catch	ladi	修改2: 去掉e
	lady	修改3: 把i变成y

ly副词后缀;
ed动词过去分词

这样，在修改2的时候，就可以找到catch，在修改3的时候就可以找到lady。

■ 汉语词法分析

特点：找出词素简单，切分出词困难。

例如：优秀人才学人才学

1. 优秀人 - 才学人才学
2. 优秀人才 - 学人才学

11.3 句法分析

- 句法分析也是自然语言处理中的基础性工作，它分析句子的句法结构（主谓宾结构）和词汇间的依存关系（并列，从属等）。
 - 通过句法分析，可以为语义分析，情感倾向，观点抽取等NLP应用场景打下坚实的基础。
- 最常见的句法分析任务：
 1. 句法结构分析，识别句子的主谓宾定状补，并分析各成分之间的关系。
 2. 语义依存关系分析，识别词汇间的从属、并列、递进等关系，可以获得较深层的语义信息。



11.4 语义分析

- 语义分析是将句法成分与应用领域中的目标表示相关联。
 1. 文本分类
 2. 情感分析
 3. 意图识别

11.5 词库

- WordNet词库：1990年由Princeton大学的Miller等人设计和构造的。
- 包含将近95600个词形(51500单词和44100搭配词)和70100个词义，分为名词、动词、形容词、副词和虚词5类。
- WordNet中，按语义而不是按词性来组织词汇信息，名词有57000个，含有48800个同义词集，分成25类文件，平均深度12层。最高层为根概念，不含有固有名词。

● 中文分词库

分词工具	词库中词的数量	最后更新时间
jieba	16.6万	2015年
IK	27.5万	2012年
mmseg	15万	2014年
word	64.2万	2015年

11.5 词库

● 中文分词方法

- (1) **最大匹配法**: 在计算机中存放一个分词用词典, 从待切分的文本中按自左到右的顺序截取一个定长的汉字串, 与词典中的词进行匹配, 若匹配不成功, 则把该字符串从右边逐次减去一个汉字, 再与词典中的词进行匹配, 直到成功为止。
- (2) **逆向最大匹配法**: 从待切分文本中截取字符串的方向是从右到左。匹配不成功时, 将所截取的汉字串从左至右逐次减去一个汉字, 再与词典中的词进行匹配, 直到匹配成功为止。
- (3) **逐词遍历匹配法**: 逐词遍历匹配法中存放的词按由长到短的顺序, 逐个与待切分的语料文本进行匹配, 直到把文本中的所有词都切分出来为止。

11.5 词库

- Jieba分词支持三种分词模式
- 精确模式：试图将句子最精确地切开，适合文本分析；
- 全模式：把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；
- 搜索引擎模式：在精确模式的基础上，对长词再次切分，适合用于搜索引擎分词。
- 另外它还自带有关键词摘取算法。



11.6 机器翻译

1946年，沃伦·韦弗提出了用机器将一种语言翻译成另一种语言

- 基于规则的机器翻译
- 基于实例的机器翻译
- 基于统计的机器翻译
- 神经机器翻译

11.6.1 基于规则的机器翻译

● 乔姆斯基的形式文法: $G = (T, N, S, P)$

T : 终结符集合

N : 非终结符集合

S : 起始符

P : 产生式规则集

例: $G = (T, N, S, P)$

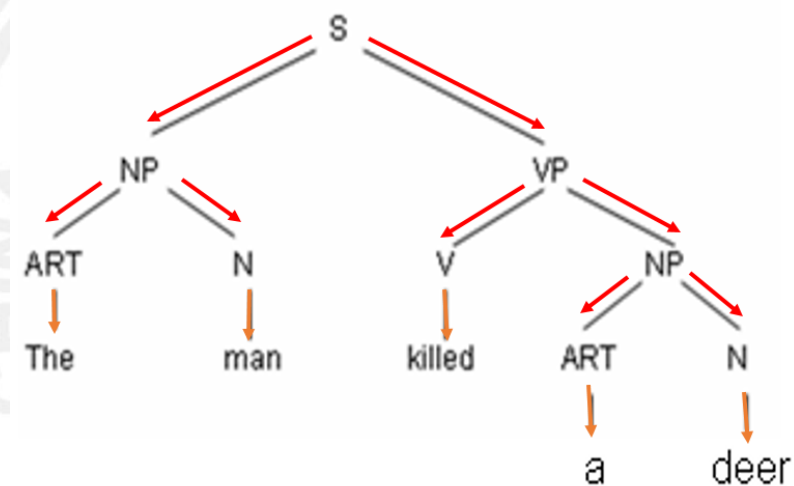
■ $T = (\text{the}, \text{man}, \text{killed}, \text{a}, \text{deer})$

■ $N = (S, NP, VP, N, ART, V)$

■ $S = S$

■ P : (1) $S \rightarrow NP + VP$ (2) $NP \rightarrow ART + N$ (3) $VP \rightarrow V + NP$

(4) $ART \rightarrow \text{the} | \text{a}$ (5) $N \rightarrow \text{man} | \text{deer}$ (6) $V \rightarrow \text{killed}$



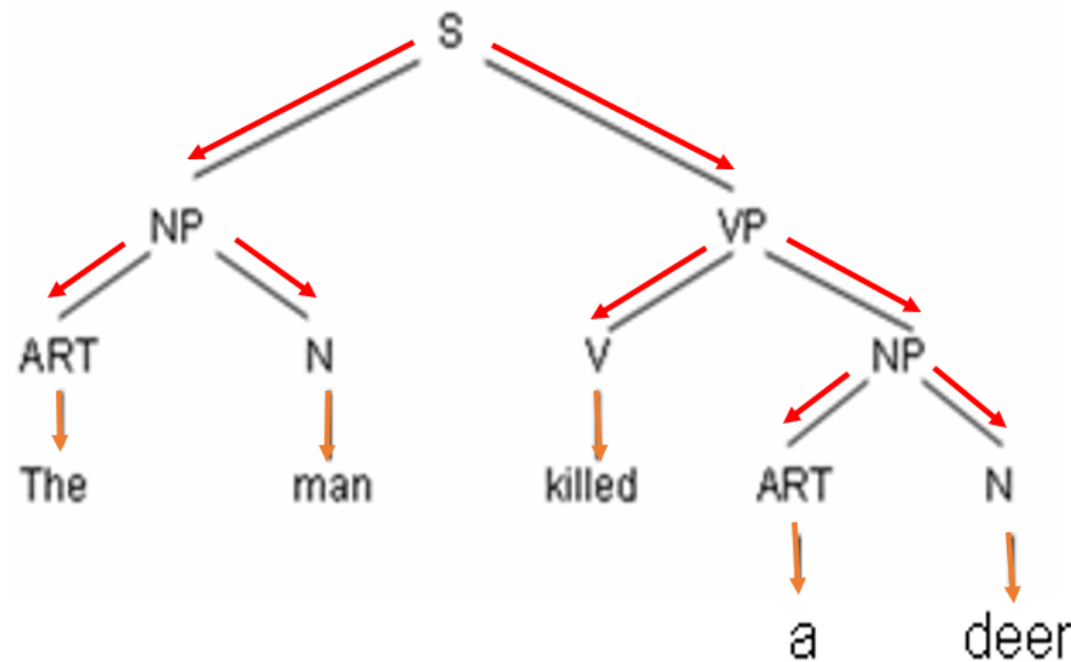
句法分析树举例

11.6.1 基于规则的机器翻译

■ 在对一个句子进行分析的过程中，如果把分析句子各成分间关系的推导过程用树形图表示出来，那么这种图称为句法分析树。

■ 例如：The man killed a deer.

$S \rightarrow NP + VP \quad \rightarrow ART + N + VP$
 $\rightarrow The\ man + VP \quad \rightarrow The\ man + V + NP$
 $\rightarrow The\ man\ killed + NP$
 $\rightarrow The\ man\ killed + ART + N$
 $\rightarrow The\ man\ killed\ a\ deer.$



句法分析树举例

11.6.1 基于规则的机器翻译

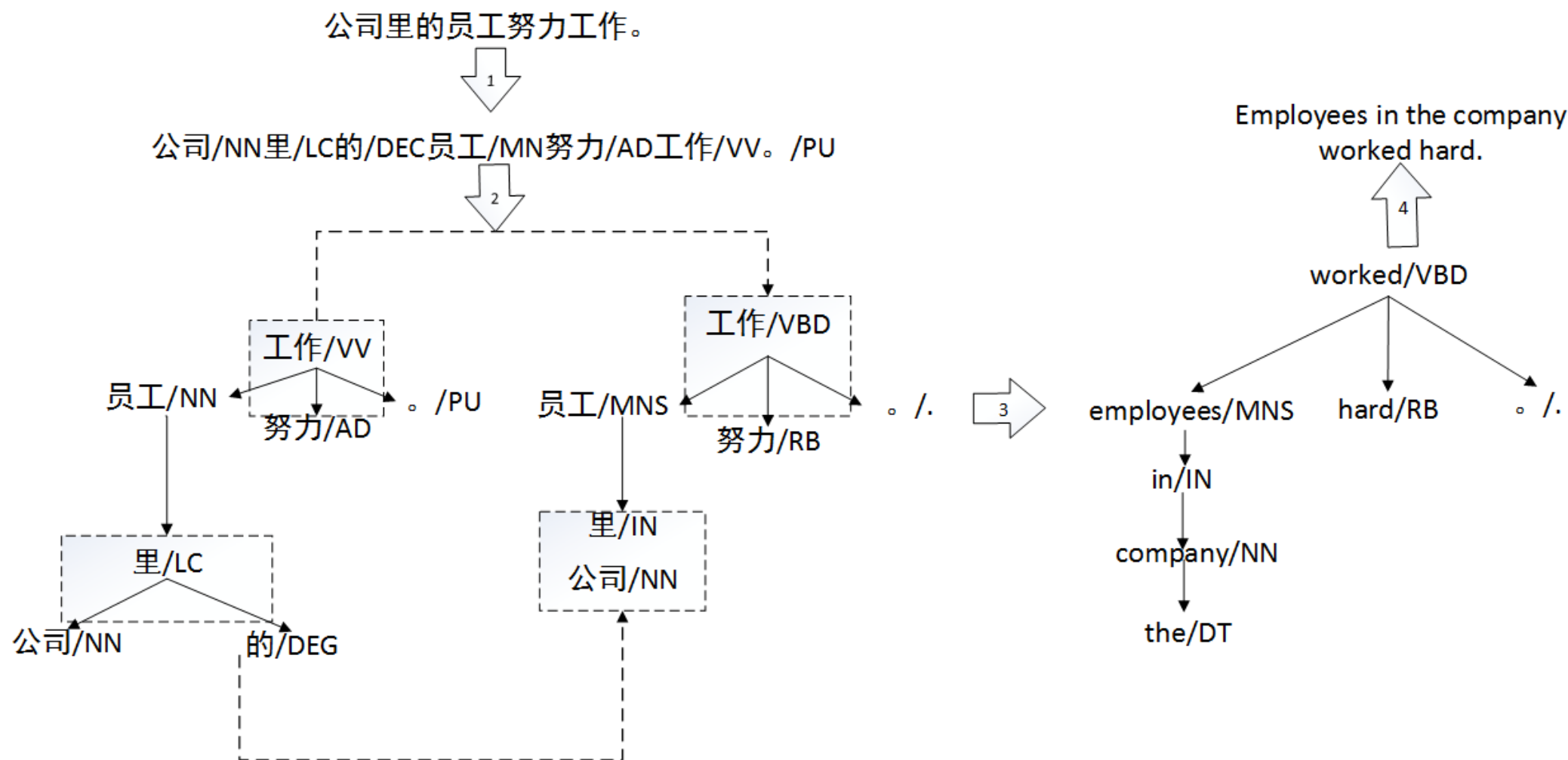


图11.4 基于规则的机器翻译流程图

11.6.2 基于实例的机器翻译

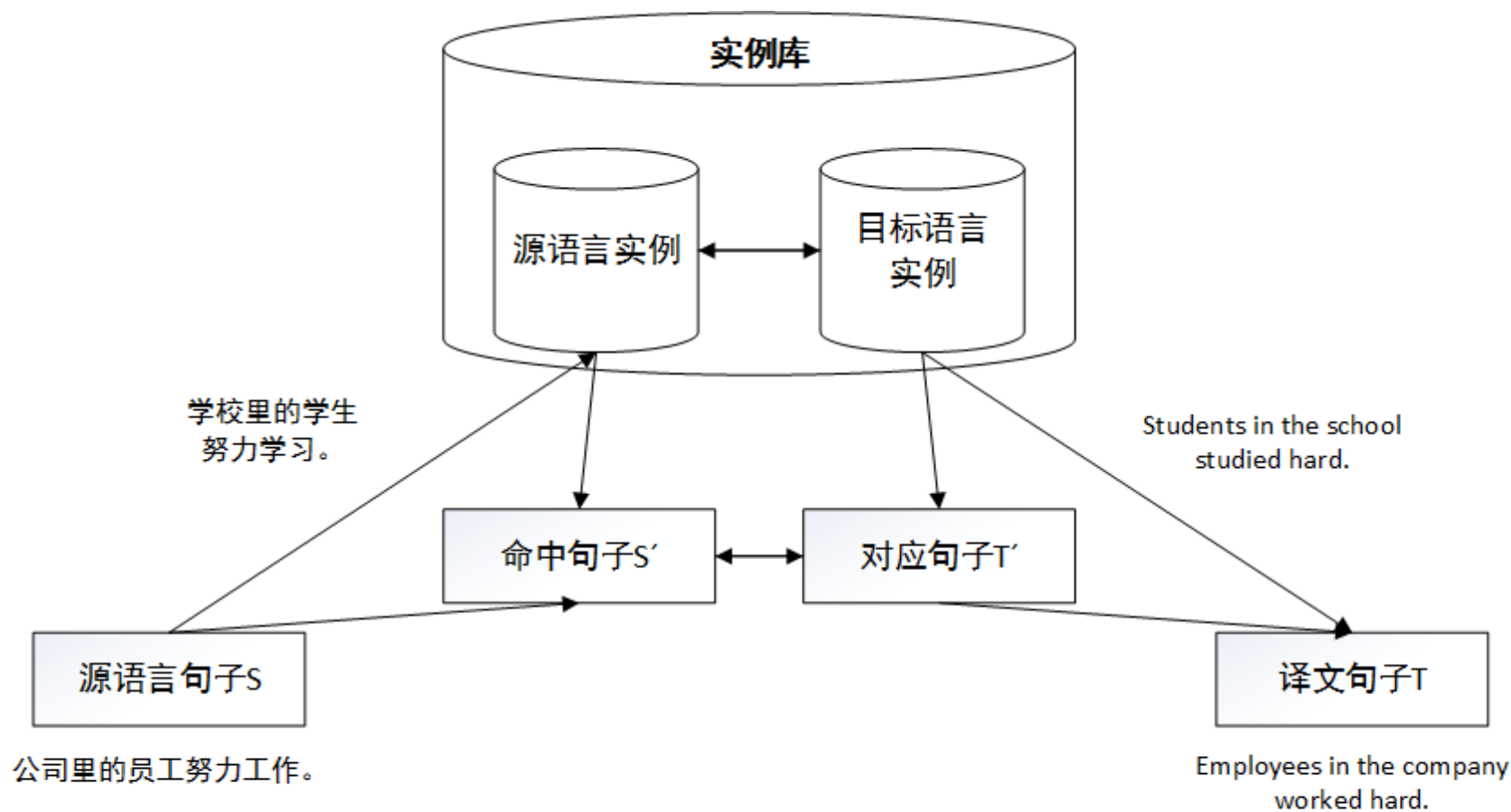
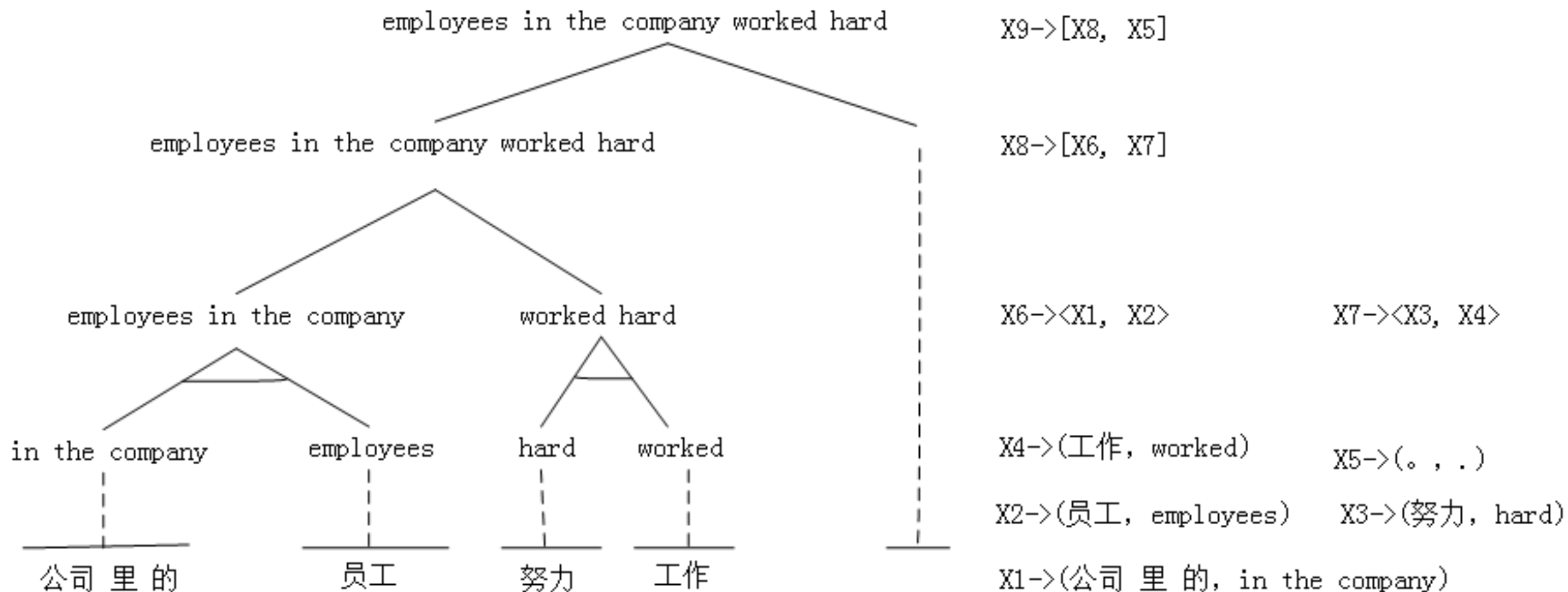


图11.5 基于实例的机器翻译流程图



11.6.3 基于统计的机器翻译



基于CKY解码算法的机器翻译实例：片段+拼接

11.6.4 神经机器翻译

中文语句→语义理解→英文语句

- 编码：把输入的源语言句子用一系列内部状态来表示
- 解码：根据语义分析结果逐词生成目标语言
- 比统计机器翻译效果好很多

中文：我就是这样评价今年的女子奥运体操队的，原因不止一个。

统计机器翻译： **I said of** this year's women's Olympic gymnastics team, because more than one.

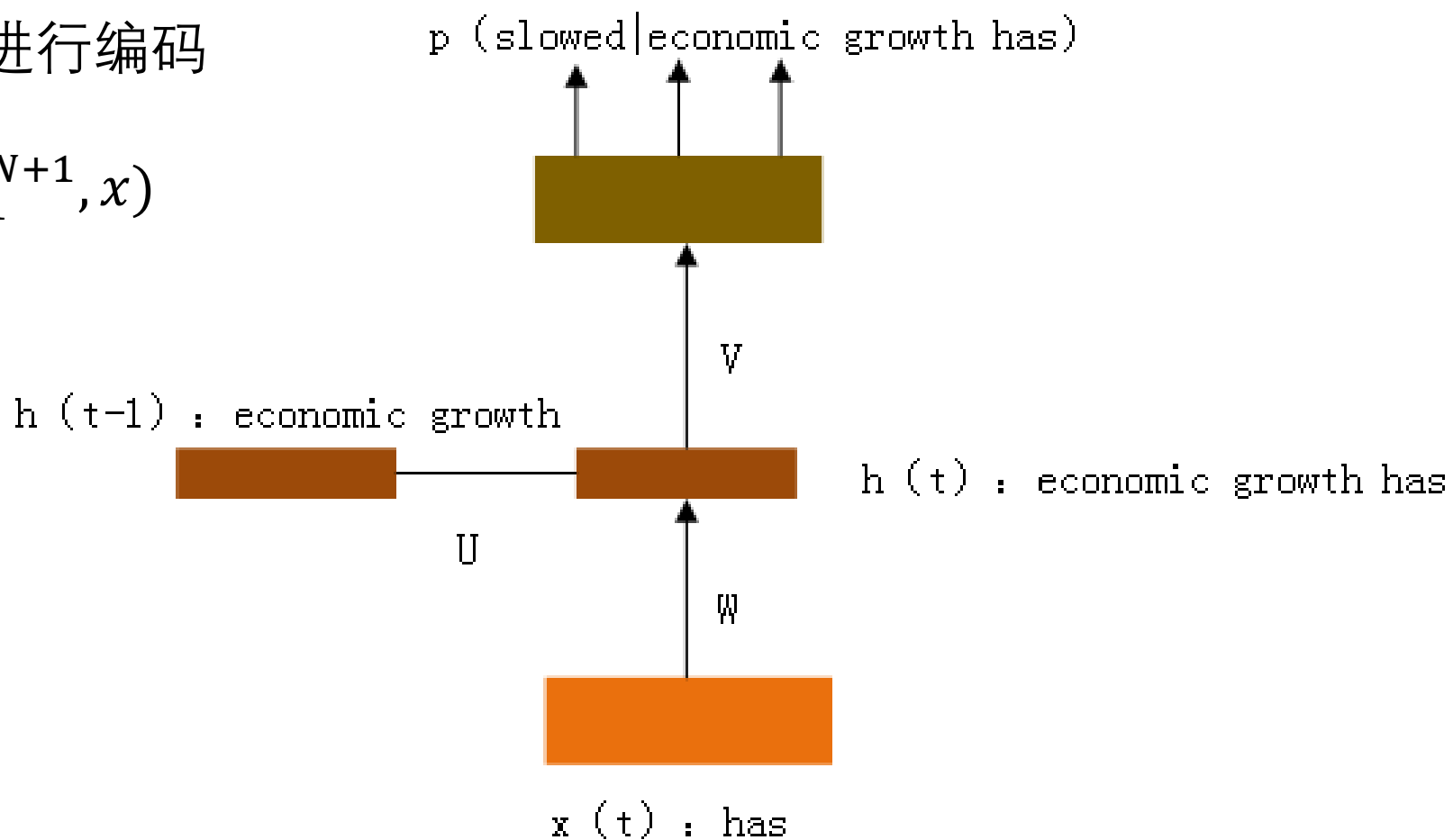
神经机器翻译： **This is how I evaluate** this year's women's Olympic gymnastics team, for more than one.

人工翻译： **That's what I call** this year's women's Olympic gymnastics team, for more than one.

11.6.4 神经机器翻译

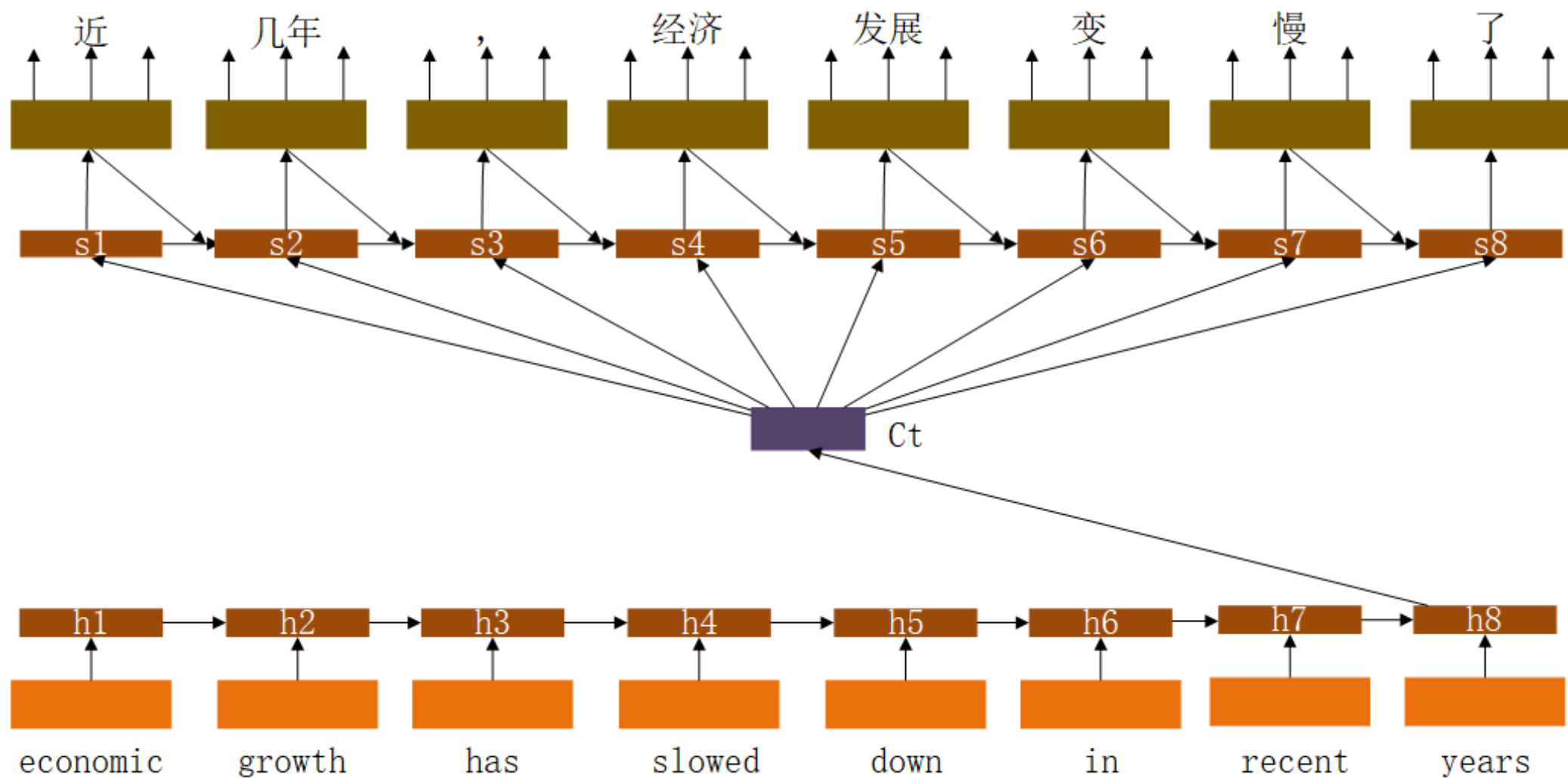
- 采用循环神经网络对句子进行编码

$$p(y_i|x) = \prod_{i=1}^{|y|} p(y_i|y_{i-1}^{i-N+1}, x)$$



编码器解码器翻译模型

11.6.4 神经机器翻译



基于源语言句子编码表示的循环神经网络翻译模型



11.7 自然语言人机交互

● 智能人机交互

- 利用自然语言实现人与机器的交流

● 应用实例：

- 面向业务的机器人—微软小娜
- 聊天机器人—微软小冰

● 智能人机交互的功能

- ✓ 通用聊天：沟通技巧、通用聊天数据、主题聊天数据，用户兴趣
- ✓ 信息服务和问答：搜索的能力、问答的能力

11.7.1 对话系统

- 以完成特定任务为主要目的的人机交互系统

- 例如：

- 智能个人助手（Apple Siri, Google Assistant, Microsoft Cortana, Facebook M等）
- 智能音箱（Amazon Alexa, Google Home, Apple Home Pod）

- 对话系统的3个模块：

- 对话理解
- 对话管理
- 回复生成

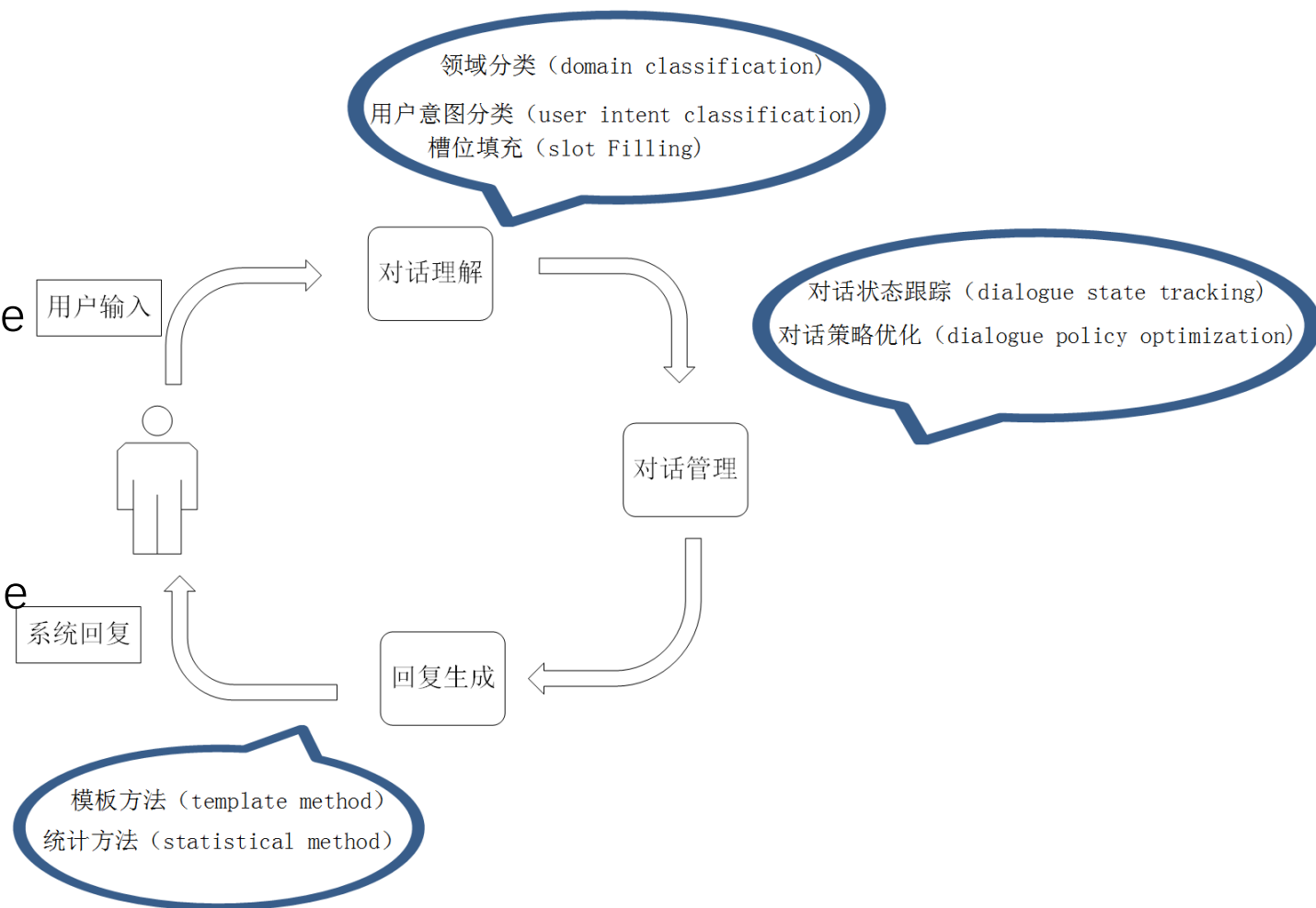


图11.11 对话系统流程图

11.7.1 对话系统

● 对话理解

- 领域分类：根据用户对话内容确定所属领域
- 用户意图分类：根据领域分类结果确定用户意图
- ✓ 分类采用基于统计的学习模型，
DBN,RNN,CNN
- 槽位填充：针对某个具体任务，抽取槽位信息
- ✓ 序列标注任务：CRF,RNN,编码器-解码器，多任务学习

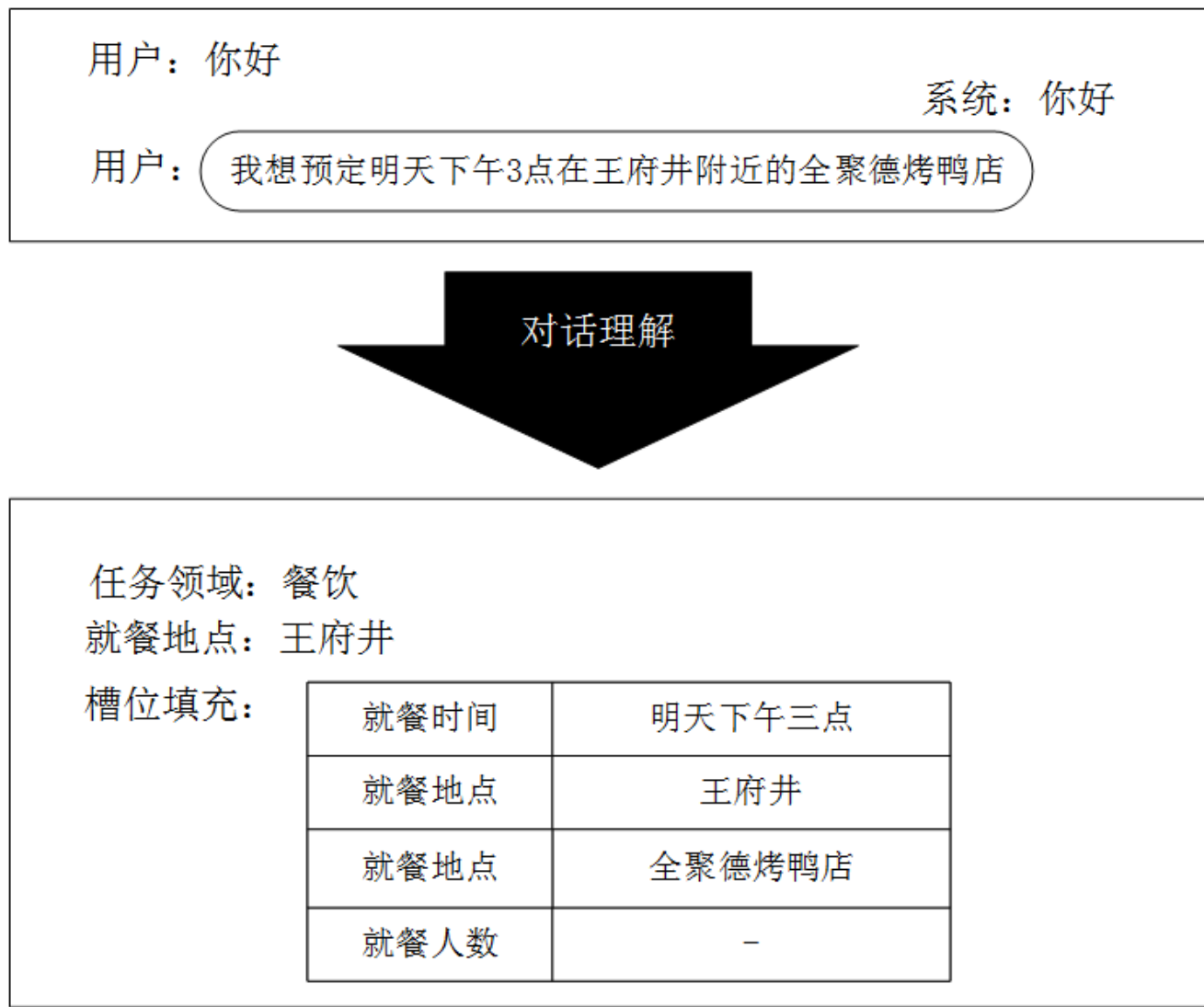


图11.12对话理解示例图

11.7.1 对话系统

● 对话管理

- 对话状态跟踪：对话结束时对整个对话状态的更新
- 对话策略优化：对话状态更新后接下来系统的行动

● 典型对话管理方法

❖ 基于有限状态机的方法

- 有明确清晰的状态转移图
- 难以应对异常情况

❖ 基于部分可观测马尔可夫决策过程

- 将不确定性引入模型，健壮性强
- 通过最大回报函数的期望进行优化

❖ 基于深度学习的方法

- 对整个对话历史进行编码，
- 采用增强学习，通过最大化未来回报进行优化

任务领域：餐饮
就餐地点：王府井

槽位填充：

就餐时间	明天下午三点
就餐地点	王府井
就餐地点	全聚德烤鸭店
就餐人数	-

对话管理

对话状态跟踪：

就餐时间	明天下午三点	0.90
就餐地点	王府井	0.85
餐厅名称	全聚德烤鸭店	0.90
就餐人数	-	0.10

对话策略优化：询问就餐人数（）

11.7.1 对话系统

● 回复生成

• 基于模板的方法

- 使用规则模板完成从系统行动指令到自然语言回复
- 模板扩展性和句子多样性不足

• 基于统计的方法

• 基于规划的方法

- 句子规划将系统行动指令转化为中间结构
- 表层实现将中间结构转换为自然语言

• 基于语料的方法

- 由行动指令直接生成自然语言句子

• 基于深度学习的方法

- 采用循环神经网络的序列生成模型

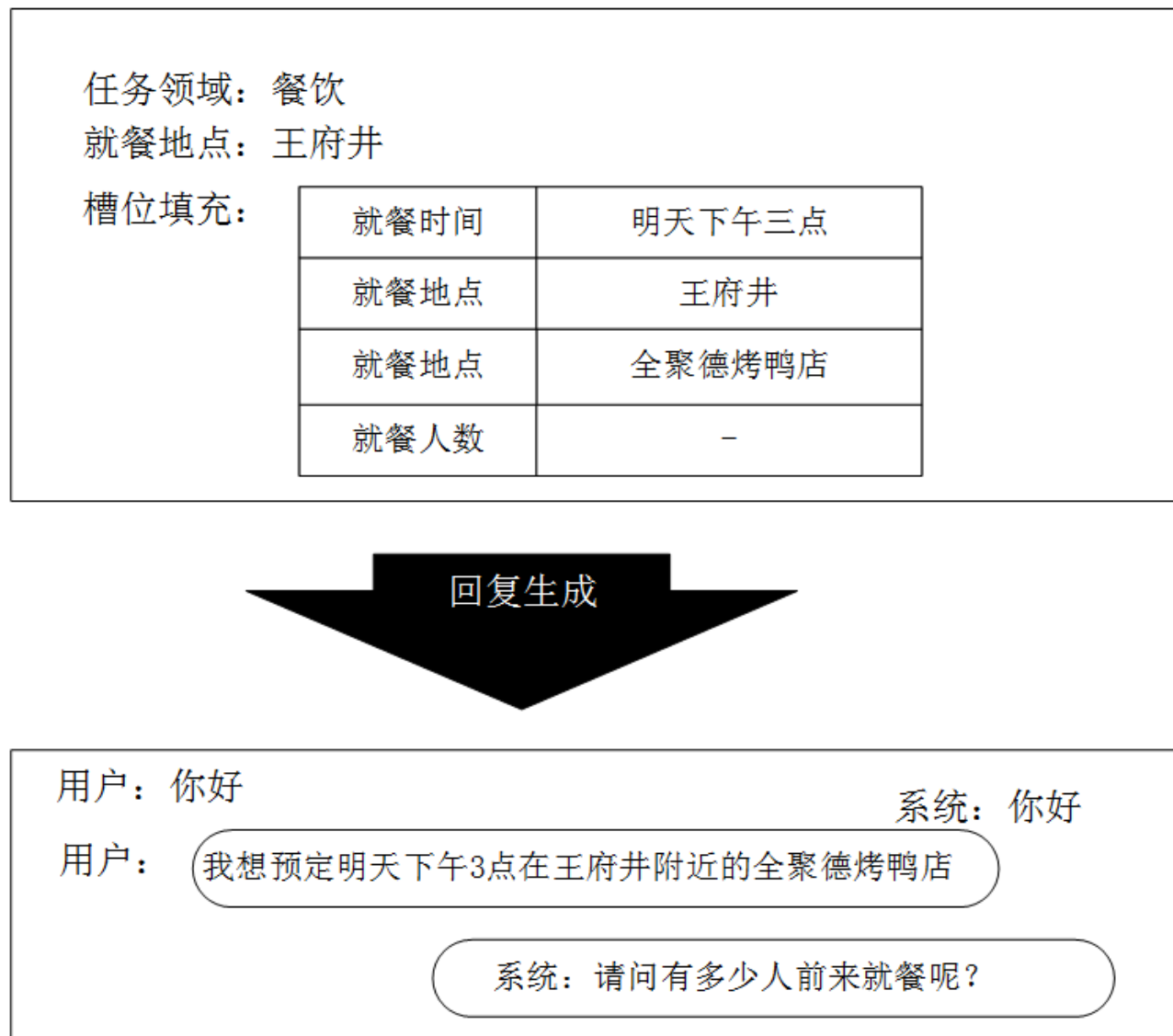


图11.14回复生成示例图

11.7.2 聊天机器人

● 基于规则的聊天机器人

预先定义一系列规则：

- 关键词回复词典
- 条件终止判断
- 输入分类器

- 1) 规则系统对输入的自然语言解析；
- 2) 抽取预定义的关键词；
- 3) 通过预先定义的模板进行回复；
- 4) 如果不在规则体系内，则用万能回复。

角色	对话内容
人	Hello.
聊天机器人	Hi. How are you?
人	I am fine.
聊天机器人	I am happy to hear that.
人	I am eating Meatball Marinara.
聊天机器人	Ok. Good for you.

11.7.2 聊天机器人

● 基于检索的聊天机器人

- 搜索引擎技术+人类对话语料
- 组织成一问一答结构
- 匹配模型是关键

与检索式聊天机器人对话历史

角色	对话内容
人	你最近怎么样？
聊天机器人	挺好的，好久不见呀。
人	你吃午饭了吗？
聊天机器人	我吃了番茄炒西红柿和马铃薯炒土豆丝，味道有点怪。
人	番茄炒西红柿是什么？
聊天机器人	胡萝卜。

11.7.2 聊天机器人

● 基于检索的聊天机器人

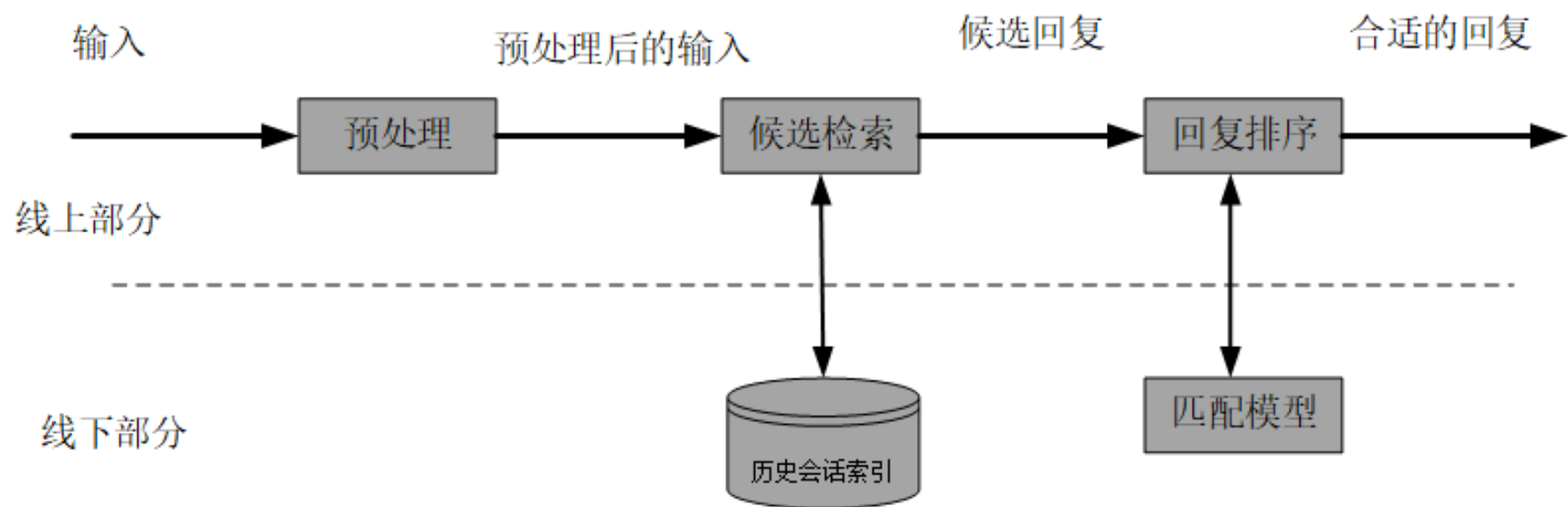


图 11.15 检索式聊天机器人架构

11.7.2 聊天机器人

- 基于生成的聊天机器人
- 序列到序列

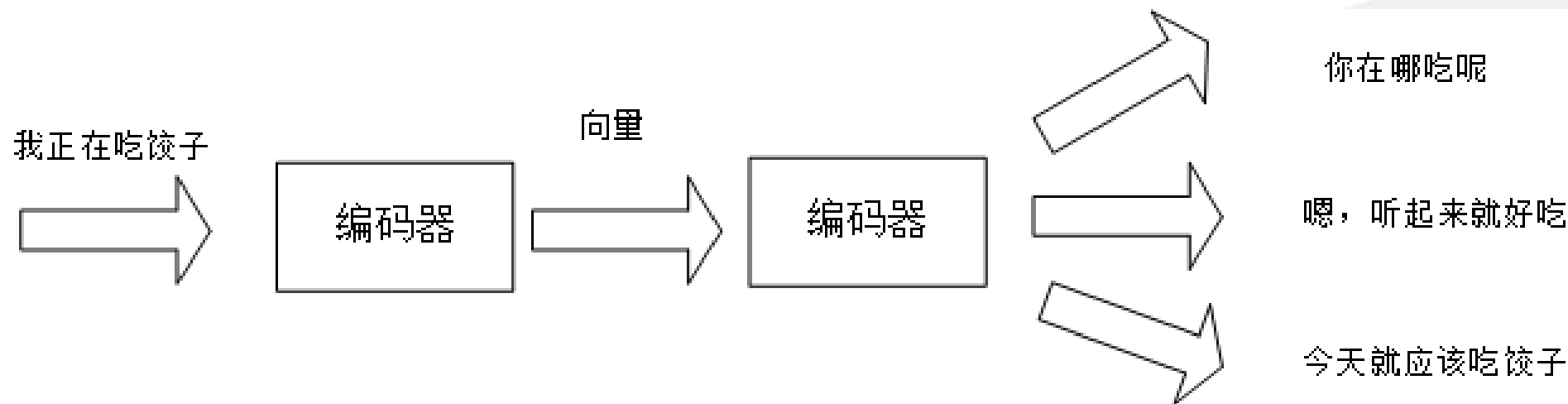


图 11.16 生成式聊天机器人序列到序列模型示意图

11.8 智能问答

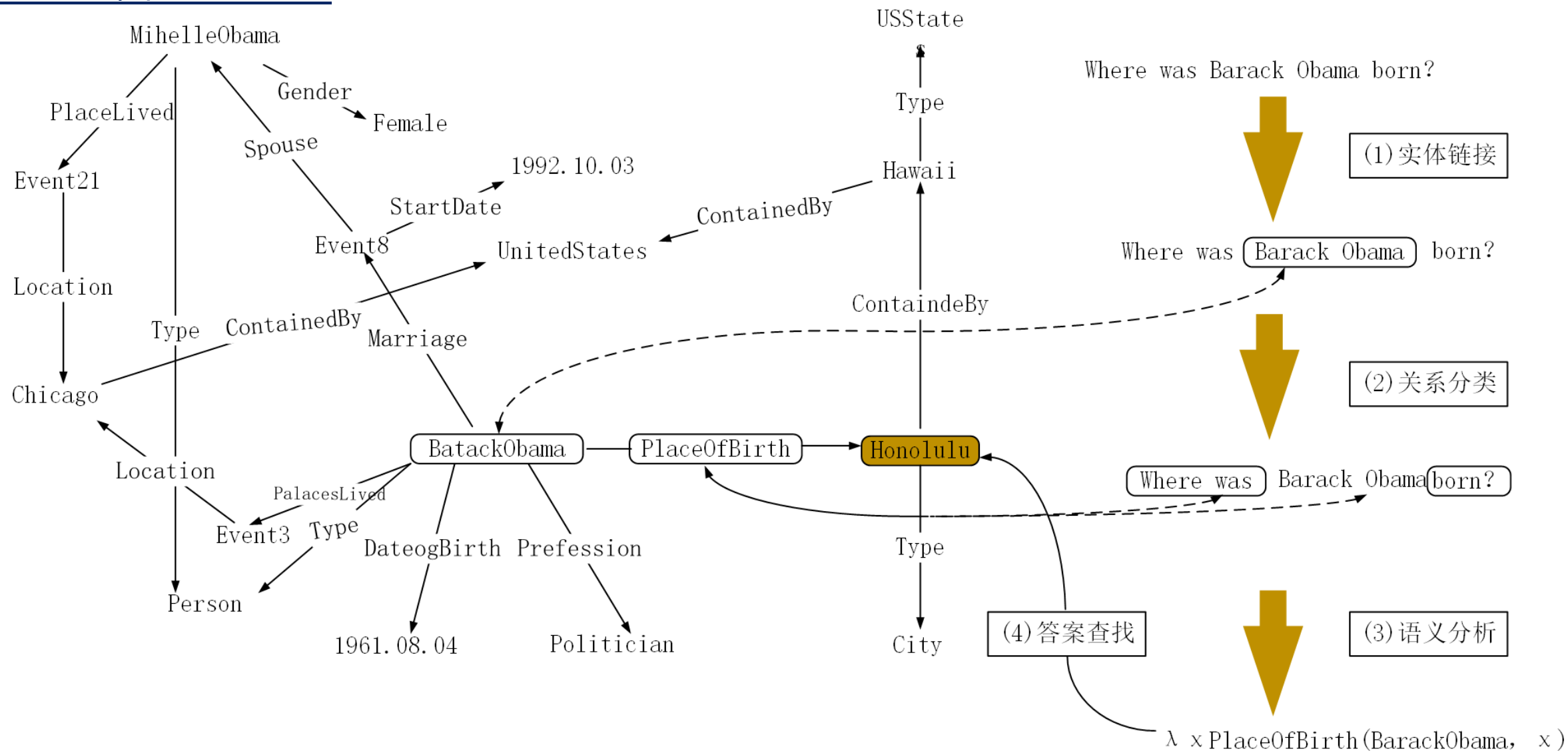


图 11.17 知识图谱问答示例

11.8 智能问答

● 基于语义分析的知识问答

• 基于文法的方法

- ✓ 抽取语义分析规则集合
- ✓ 采用基于动态规划的解析算法生成语义表示候选集合
- ✓ 基于标注数据集训练语义候选排序模型
- ✓ 返回最高分的语义作为最终结果

• 基于深度学习的方法

- ✓ 序列生成任务，采用编码器-解码器完成自然语言到语义的转换

● 基于答案排序的知识问答

- 1) 问题实体识别
- 2) 答案候选检索
- 3) 答案候选表示
- 4) 答案候选排序



THANKS

