



# Computer Version



## Natural Language for Communication



### Capacity to Process Natural Language

为什么处理自然语言 (Natural Language Processing, NLP) 的能力重要?

1. **获取信息 (Acquire information)**: NLP使计算机能够从书面语言中提取信息。例如, 你可以通过搜索引擎获取论文的核心信息, 也可以用智能助手如Siri从互联网查询天气信息。
2. **与人类交流 (Communicate with humans)**: 让计算机能够通过自然语言与人类交流, 例如聊天机器人能够用普通的对话方式回答你的问题, 这在客服和教育领域尤其有用。

为了实现自然语言的深度理解, 需要用到以下语法模型 (Grammatical models):

1. **词汇类别 (Lexical category)**: 这是关于单词的词性, 例如:
  - 名词 (Noun): 代表一个物体, 如 “dog”。
  - 形容词 (Adjective): 描述物体特征, 如 “beautiful”。
  - 动词 (Verb): 描述动作, 如 “run”。
2. **句法类别 (Syntactic category)**: 通过组合词汇类别形成更大的结构, 例如:
  - 名词短语 (Noun phrase): 由名词或修饰它的形容词组成, 例如 “a beautiful dog”。
  - 动词短语 (Verb phrase): 描述动作及其对象, 例如 “is running fast”。
3. **短语结构 (Phrase structure)**: 将句法类别组织成树状结构 (树结构是一种嵌套的表示法), 用于表示完整的句子。这有助于解析语义关系。例如:
  - “The cat sleeps on the mat” 的句子结构可以分解为主语 (The cat) + 动词短语 (sleeps on the mat)。

📖 举个很简单的例子: 假设我们让一个智能客服回答 “天气怎么样?” 这个问题。它的处理过程可能如下:

- **词汇分析**: 识别 “天气” 为名词, “怎么样” 为疑问词。
- **句法分析**: 确定整个句子是在询问天气状态。
- **短语结构**: 将句子结构化, 以便识别主语是 “天气”, 谓语是询问语气。

通过这样的分解, 计算机就能够更准确地理解并回答 “今天晴天, 温度20°C” 这样自然的语言。



### Probabilistic Context-Free Grammar (PCFG)



#### Grammar

语法 (Grammar) 是定义一种语言规则的集合, 描述一组允许的单词序列 (字符串)。

## Context-Free Grammars (CFGs)

上下文无关语法 (CFGs) 是由一组生成规则 (Production rules) 组成, 每条规则的形式是:

$$A \rightarrow \alpha$$

- $A$ : 单个非终结符号 (Non-terminal symbol), 如句子的结构。
- $\alpha$ : 由终结符号 (Terminal symbol, 例如实际单词) 或非终结符号组成的字符串。

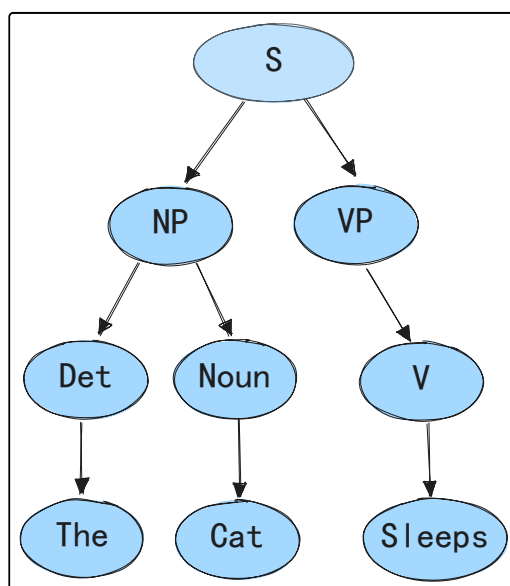
这些规则允许将非终结符替换为右侧的结构, 应用在任何上下文中。

### 实际例子

假设我们想解析句子 “猫在睡觉” (The cat sleeps), 一个可能的CFG规则是:

- $S \rightarrow NP, VP$ : 句子 ( $S$ ) 由名词短语 ( $NP$ ) 和动词短语 ( $VP$ ) 组成。
- $NP \rightarrow Det, N$ : 名词短语由限定词 ( $Det$ ) 和名词 ( $N$ ) 组成。
- $VP \rightarrow V$ : 动词短语由动词 ( $V$ ) 组成。

对于 “猫在睡觉”, 解析树可能是:



## What is PCFG?

PCFG在CFG的基础上为每条生成规则分配概率, 这些概率用于描述字符串生成的可能性。

例如:

- $VP \rightarrow Verb, [0.70]$ : 动词短语中动词的生成概率为70%。
- $VP \rightarrow VP, NP, [0.30]$ : 动词短语中包含名词短语的概率为30%。

这些概率的总和为1, 可以帮助计算解析树的可能性。

### 实际例子

假设我们有以下PCFG规则:

- $S \rightarrow NP, VP, [1.0]$
- $NP \rightarrow Det, N, [0.8]$
- $VP \rightarrow Verb, [0.7]$
- $VP \rightarrow VP, NP, [0.3]$

句子 “The cat sleeps” 可能生成的概率是:

$$P(S \rightarrow NP, VP) = 1.0$$

$$2. P(NP \rightarrow Det, N) = 0.8$$

$$3. P(VP \rightarrow Verb) = 0.7$$

$$\text{总概率为: } P(\text{Tree}) = 1.0 \times 0.8 \times 0.7 = 0.56$$

这表示生成该句子的可能性为56%。

A Toy Language  $\mathcal{E}_0$

## 🍷 Lexical Categories

词汇类别将单词分为不同的语法功能组，例如：

<i>Noun</i>	$\rightarrow$	<b>stench</b> [0.05]   <b>breeze</b> [0.10]   <b>wumpus</b> [0.15]   <b>pits</b> [0.05]   ...
<i>Verb</i>	$\rightarrow$	<b>is</b> [0.10]   <b>feel</b> [0.10]   <b>smells</b> [0.10]   <b>stinks</b> [0.05]   ...
<i>Adjective</i>	$\rightarrow$	<b>right</b> [0.10]   <b>dead</b> [0.05]   <b>smelly</b> [0.02]   <b>breezy</b> [0.02] ...
<i>Adverb</i>	$\rightarrow$	<b>here</b> [0.05]   <b>ahead</b> [0.05]   <b>nearby</b> [0.02]   ...
<i>Pronoun</i>	$\rightarrow$	<b>me</b> [0.10]   <b>you</b> [0.03]   <b>I</b> [0.10]   <b>it</b> [0.10]   ...
<i>RelPro</i>	$\rightarrow$	<b>that</b> [0.40]   <b>which</b> [0.15]   <b>who</b> [0.20]   <b>whom</b> [0.02] $\vee$ ...
<i>Name</i>	$\rightarrow$	<b>John</b> [0.01]   <b>Mary</b> [0.01]   <b>Boston</b> [0.01]   ...
<i>Article</i>	$\rightarrow$	<b>the</b> [0.40]   <b>a</b> [0.30]   <b>an</b> [0.10]   <b>every</b> [0.05]   ...
<i>Prep</i>	$\rightarrow$	<b>to</b> [0.20]   <b>in</b> [0.10]   <b>on</b> [0.05]   <b>near</b> [0.10]   ...
<i>Conj</i>	$\rightarrow$	<b>and</b> [0.50]   <b>or</b> [0.10]   <b>but</b> [0.20]   <b>yet</b> [0.02] $\vee$ ...
<i>Digit</i>	$\rightarrow$	<b>0</b> [0.20]   <b>1</b> [0.20]   <b>2</b> [0.20]   <b>3</b> [0.20]   <b>4</b> [0.20]   ...

每一类别中，各选项的概率之和为 1。这确保了每次生成句子时，能够根据概率选出最有可能的单词。

## 🍷 Syntactic Categories

句法类别定义了句子结构和短语的生成规则。例如：

$S$	$\rightarrow$	$NP\ VP$	[0.90]	I + feel a breeze
	$ $	$S\ Conj\ S$	[0.10]	I feel a breeze + and + It stinks
$NP$	$\rightarrow$	$Pronoun$	[0.30]	I
	$ $	$Name$	[0.10]	John
	$ $	$Noun$	[0.10]	pits
	$ $	$Article\ Noun$	[0.25]	the + wumpus
	$ $	$Article\ Adjs\ Noun$	[0.05]	the + smelly dead + wumpus
	$ $	$Digit\ Digit$	[0.05]	3 4
	$ $	$NP\ PP$	[0.10]	the wumpus + in 1 3
	$ $	$NP\ RelClause$	[0.05]	the wumpus + that is smelly
$VP$	$\rightarrow$	$Verb$	[0.40]	stinks
	$ $	$VP\ NP$	[0.35]	feel + a breeze
	$ $	$VP\ Adjective$	[0.05]	smells + dead
	$ $	$VP\ PP$	[0.10]	is + in 1 3
	$ $	$VP\ Adverb$	[0.10]	go + ahead
$Adjs$	$\rightarrow$	$Adjective$	[0.80]	smelly
	$ $	$Adjective\ Adjs$	[0.20]	smelly + dead
$PP$	$\rightarrow$	$Prep\ NP$	[1.00]	to + the east
$RelClause$	$\rightarrow$	$RelPro\ VP$	[1.00]	that + is smelly

句子结构 (Sentence Structure) :  $S \rightarrow NP, VP, [0.90], |, S, Conj, S, [0.10]$

第一条规则表示句子由名词短语 ( $NP$ ) 和动词短语 ( $VP$ ) 组成，概率为 0.90。

第二条规则表示句子也可以由两个句子通过连接词 ( $Conj$ ) 连接而成，概率为 0.10。

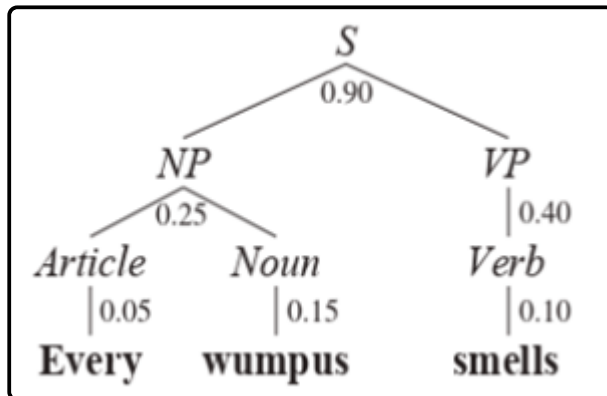
名词短语 (Noun Phrase, NP) :

$NP \rightarrow Pronoun, [0.30], |, Name, [0.10], |, Article, Noun, [0.25], |, \dots$

动词短语 (Verb Phrase, VP) :  $VP \rightarrow Verb, [0.40], |, VP, NP, [0.35], |, \dots$

## 🔥 Phrase Structure Example

句子 “Every wumpus smells” 的短语结构解析如下：



- $S \rightarrow NP, VP, [0.90]$
- $NP \rightarrow Article, Noun, [0.25]$ 
  - $Article \rightarrow Every, [0.05]$
  - $Noun \rightarrow wumpus, [0.15]$
- $VP \rightarrow Verb, [0.40]$ 
  - $Verb \rightarrow smells, [0.10]$

整棵树的概率为：  $P(\text{Tree}) = 0.90 \times 0.25 \times 0.05 \times 0.15 \times 0.40 \times 0.10 = 0.000675$

🔔 通过这个例子，我们看到生成句子的概率依赖于所有规则的联合概率。这种方式可以帮助模型评估不同句子的合理性，从而选择最优的句子结构。

## 🔥 Syntactic Parsing

### 🔥 Definition

句法解析 (Syntactic Parsing) 是根据语法规则，解析一串单词的短语结构的过程。可以通过两种方法完成：

1. **Top-Down** (从上到下)：从句子的最高层结构 ( $S$ ) 开始，逐步解析其组成部分。
2. **Bottom-Up** (从下到上)：从单词 (终结符号) 开始，逐步构建到完整的句子。

🖼️ Example: Parsing “The wumpus is dead”

解 析 过 程 如 下：

List of items	Rule
$S$	
$NP VP$	$S \rightarrow NP VP$
$NP VP Adjective$	$VP \rightarrow VP Adjective$
$NP Verb Adjective$	$VP \rightarrow Verb$
$NP Verb dead$	$Adjective \rightarrow dead$
$NP is dead$	$Verb \rightarrow is$
$Article Noun is dead$	$NP \rightarrow Article Noun$
$Article wumpus is dead$	$Noun \rightarrow wumpus$
$the wumpus is dead$	$Article \rightarrow the$

1. 应用规则  $S \rightarrow NP, VP$ ：句子分解为名词短语 ( $NP$ ) 和动词短语 ( $VP$ )。
2.  $NP$  进一步解析为  $Article, Noun$ ，如 “The wumpus”。
3.  $VP$  解析为  $Verb, Adjective$ ，如 “is dead”。

最终的解析结构为：

- $S \rightarrow NP, VP$
  - $NP \rightarrow Article, Noun$  (The wumpus)
  - $VP \rightarrow Verb, Adjective$  (is dead)
- 

## Ambiguity in Parsing

### Intended Meaning

有些句子对人类来说毫不含糊，但对机器而言却很模糊。例如：

- "Squad helps dog bite victim."
- "Include your children when baking cookies."
- "Milk drinkers are turning to powder."

这些句子需要机器从语义或上下文中理解其真正意图。

### Types of Ambiguity

#### 1. Lexical Ambiguity (词汇歧义)

- 一个单词有多个含义，例如 "bank" 既可以指银行，也可以指河岸。

#### 2. Syntactic Ambiguity (句法歧义)


- 一个短语有多种解析方式。例如：
  - "I smelled a wumpus in 2.2" :
    - 解释1: "in 2.2" 修饰 "wumpus" .
    - 解释2: "in 2.2" 修饰 "smelled" .

#### 3. Semantic Ambiguity (语义歧义)

- 同一句话可以有多种含义。例如：
  - "I saw her duck" :
    - 解释1: 我看到了她的鸭子。
    - 解释2: 我看到她低下头。

#### 4. Metonymy (转喻)

- 一种修辞方式，用一个对象表示另一个对象。例如：
    - "Chrysler announced a new model" :
      - 这里 "Chrysler" 实际上指代 "Chrysler公司" .
- 

 总结：句法解析是NLP的关键步骤，用于构建句子的结构和语义关系。处理模糊性 (ambiguity) 是解析中的主要挑战，而结合语义规则和上下文信息可以有效降低歧义。



# Disambiguation



## Definition

歧义消解 (Disambiguation) 是通过一定的概率模型来解释句子中的模糊之处。然而, 这些概率模型通常代表的是一般知识, 而不是特定场景。为了更准确地消解歧义, 我们需要结合以下模型:

### 1. World Model (世界模型):

- 描述某种事件在现实世界中发生的可能性。
- 例子: “I am dead” 可能是某个角色在电影中的台词, 也可能是某人表达惊讶的比喻。

### 2. Mental Model (心理模型):

- 捕捉说话者试图向听众传达的意图。
- 例子: “I am not a crook” 可能表达否认或强调诚实。

### 3. Language Model (语言模型):

- 描述一个单词序列被选择的可能性。
- 例子: “The quick brown fox jumps” 是自然的, 而 “Jumps quick brown fox the” 显得不符合语法规则。

### 4. Acoustic Model (声学模型):

- 处理语音沟通中音素和单词之间的映射。
- 例子: 语音助手需要判断用户说的是 “weather” 还是 “whether”。



# Recap: Natural Language Processing (NLP)



## Knowledge Acquisition (知识获取)

### Language Models (语言模型):

- 用于预测单词序列的概率, 应用于文本生成、机器翻译等任务。

### NLP Tasks:

- 包括文本分类 (Text Classification)、信息检索 (Information Retrieval)、信息抽取 (Information Extraction) 等。



## Communication (交流)

### Grammatical Models (语法模型):

- 涉及词汇类别、句法类别和短语结构。

### PCFGs, Parsing:

- 用概率上下文无关语法和解析技术分析句子结构。

### Ambiguity and Disambiguation (歧义与消解):

- 处理词汇、句法和语义层面的模糊性。