Natural Language Processing I

3. 句法分析

WU Xiaokun 吴晓堃

xkun.wu [at] gmail

2022/03/18

谁吃谁?

下面哪种写法正确?

Ich esse Fisch
Fisch esse Ich

Fisch isst mich
mich isst Fisch

– 德语练习

下面哪种写法正确?

> *Ich esse Fisch*
> *Fisch esse Ich*
>
> *Fisch isst mich*
> *mich isst Fisch*
>
> *– 德语练习*

- 只有谓语的位置固定（在主句中只有一个动词放第二位，两个以上动词、在从句中所有动词放句尾）

下面哪种写法正确?

> *Ich esse Fisch*
> *Fisch esse Ich*
>
> *Fisch isst mich*
> *mich isst Fisch*
>
> *– 德语练习*

- 只有谓语的位置固定（在主句中只有一个动词放第二位，两个以上动词、在从句中所有动词放句尾）
- 拉丁语、俄语：语序完全自由（6种组合），把强调的成分放在句首
  - 拉丁语：最常用主宾谓结构 SOV
  - 汉语：“把”、“被”

语言分类

想要研究NLP，先要了解自然语言

- **分析语 Analytic language**：語序（决定语义）、虚词（表述其他组合）
  - 汉藏语系
- **综合语 Synthetic language**：词形变化、合成
  - 屈折语：拉丁语，印欧语系
  - 黏着语：日本－琉球语系，突厥语

语言分类

想要研究NLP，先要了解自然语言

- **分析语 Analytic language**：語序（决定语义）、虚词（表述其他组合）
  - 汉藏语系
- **综合语 Synthetic language**：词形变化、合成
  - 屈折语：拉丁语，印欧语系
  - 黏着语：日本－琉球语系，突厥语

词形变化 inflections

- 我吃鱼
- I eat fish(-, es)
- Je mange [manger] du poisson(s)

屈折语

语序：Ich esse [essen] Fisch(e)

```
Ich esse Fisch
Fisch esse Ich

Fisch isst mich
mich isst Fisch
```

屈折语

语序：Ich esse [essen] Fisch(e)

```
Ich esse Fisch
Fisch esse Ich

Fisch isst mich
mich isst Fisch
```

省略：(Yo) como [comer] pescado(s)

- 西班牙语比较接近拉丁语

屈折语

语序：Ich esse [essen] Fisch(e)

```
Ich esse Fisch
Fisch esse Ich

Fisch isst mich
mich isst Fisch
```

省略：(Yo) como [comer] pescado(s)

- 西班牙语比较接近拉丁语

注意：汉语也有特殊情况

```
何辞为？--《史记·项羽本纪·鸿门宴》
给钱我。--方言
```

语法

要理解语言，就绕不开**语法 grammar**。

- 问题：语法学习很枯燥、语法书很厚
  - 被语法学习支配的痛苦
- 正确思路：语法**是工具而不是目的**

语法

要理解语言，就绕不开**语法 grammar**。

- 问题：语法学习很枯燥、语法书很厚
  - 被语法学习支配的痛苦
- 正确思路：语法**是工具而不是目的**

NLP 的最终目标：**提取语义、理解语言**。

- 简单、相对完备，但好用
- 容易实现

句法

> ***句法 Syntax**指一门语言里支配句子结构，决定词、短语、从句等句子成分如何组成其上级成分，直到组成句子的**规则或过程**。*
>
> *– Wikipedia*

句法

> ***句法 Syntax**指一门语言里支配句子结构，决定词、短语、从句等句子成分如何组成其上级成分，直到组成句子的**规则或过程**。*
>
> *– Wikipedia*

- **syntax** 来自于希腊语 *sýntaxis*：安排到一起
- 汉语句法研究主要关注语序。

语料库 corpus

部分示例取自 ATIS (Air Traffic Information System) ([Hemphill 1990]) 数据集。

- 例如：I'd like to fly to Atlanta.

构成法

构成句法

构成句法 Syntactic Constituency：不断将句子成分组合成更高一级单元的过程

- 字 -> 词 -> 短语 -> 句子。

构成句法

构成句法 Syntactic Constituency：不断将句子成分组合成更高一级单元的过程

- 字 -> 词 -> 短语 -> 句子。

名词短语 noun phrase
中国人民银行
Harry the Horse

如何分组?

合理假设：同时出现在类似的句法环境中。

- 例如：在动词之前

如何分组?

合理假设：同时出现在类似的句法环境中。

- 例如：在动词之前

名词短语在动词前
中国人民银行简称人民银行、人行或央行。
Harry the Horse is the first to pass the line.

如何分组?

合理假设：同时出现在类似的句法环境中。

- 例如：在动词之前

名词短语在动词前
中国人民银行简称人民银行、人行或央行。
Harry the Horse is the first to pass the line.

注意：非充要条件。

分组方法：词类

形容词：指代
安达尔人、洛伊拿人和先民的女王 Queen of the Andals, the Rhoynars and the First Men、七国女王/统治者 Queen/Lord of the Seven Kingdoms、全境守护 Protector of the Realm、大草海的卡丽熙 Khaleesi of the Great Grass Sea、镣铐/锁链破除者 Breaker of Shackles/Chains、弥林女王 Queen of Meereen、龙石岛公主 Princess of Dragonstone、不焚者 Unburnt、龙之母 Mother of Dragons、弥莎 Mhysa、母亲 Mother、银发女王 Silver Queen、银发女士 Silver Lady、龙女王 Dragon Queen 丹妮莉丝·坦格利安 Daenerys Targaryen。–《冰与火之歌》

分组方法：词类

形容词：指代

安达尔人、洛伊拿人和先民的女王 Queen of the Andals, the Rhoynars and the First Men、七国女王/统治者 Queen/Lord of the Seven Kingdoms、全境守护 Protector of the Realm、大草海的卡丽熙 Khaleesi of the Great Grass Sea、镣铐/锁链破除者 Breaker of Shackles/Chains、弥林女王 Queen of Meereen、龙石岛公主 Princess of Dragonstone、不焚者 Unburnt、龙之母 Mother of Dragons、弥莎 Mhysa、母亲 Mother、银发女王 Silver Queen、银发女士 Silver Lady、龙女王 Dragon Queen 丹妮莉丝·坦格利安 Daenerys Targaryen。–《冰与火之歌》

形容词：头衔

翰林学士朝散大夫右谏议大夫知制诰兼侍讲同提举万寿观公事兼判集贤院上护军河内郡开国侯食邑一千三百户赐紫金鱼袋臣 司马光 奉敕编集 –《资治通鉴》

分组方法：语序

合理假设：语序变化不影响语义。

前置 preposed、后置 postposed
On September seventeenth, I'd like to fly from Atlanta to Denver
I'd like to fly *on September seventeenth* from Atlanta to Denver
I'd like to fly from Atlanta to Denver *on September seventeenth*

分组方法：语序

合理假设：语序变化不影响语义。

前置 preposed、后置 postposed
On September seventeenth, I'd like to fly from Atlanta to Denver
I'd like to fly *on September seventeenth* from Atlanta to Denver
I'd like to fly from Atlanta to Denver *on September seventeenth*

注意：语序对中文非常重要。

语境无关语法

语境无关语法 CFG

语境无关语法 **Context-Free Grammars (CFG)**

- **规则 rules 或 生成符号 productions**
- **词典 lexicon**

名词短语 Noun Phrase (NP)

$$NP \rightarrow \text{限定词 } Det + \text{名词性成分 } Nominal$$

$$NP \rightarrow \text{专有名词 } ProperNoun$$

$$Nominal \rightarrow Noun | Nominal + Noun$$

语境无关语法 CFG

语境无关语法 **Context-Free Grammars (CFG)**

- **规则 rules 或 生成符号 productions**
- **词典 lexicon**

名词短语 Noun Phrase (NP)

$$NP \rightarrow \text{限定词 } Det + \text{名词性成分 } Nominal$$
$$NP \rightarrow \text{专有名词 } ProperNoun$$
$$Nominal \rightarrow Noun | Nominal + Noun$$

又称**短语结构文法 Phrase-Structure Grammars**

- 构成方法与**Backus-Naur Form (BNF)**等价。
- 本质上是形式化语法

层次化规则

CFG中的规则可以不断“生长”：

- 右侧：可以是生成符号、字的有序列表
- 左侧：一定是生成符号

规则细化

$$Det \rightarrow a|the$$

$$Noun \rightarrow flight$$

层次化规则

CFG中的规则可以不断“生长”：

- 右侧：可以是生成符号、字的有序列表
- 左侧：一定是生成符号

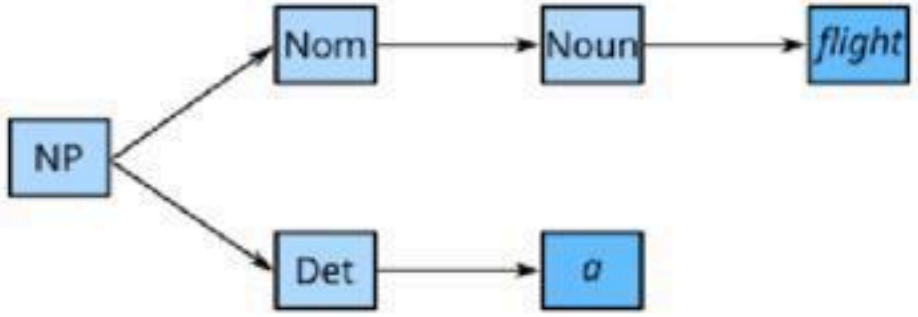
规则细化

$$Det \rightarrow a|the$$

$$Noun \rightarrow flight$$

规则导出的**层次化拓扑关系**对应于句法。

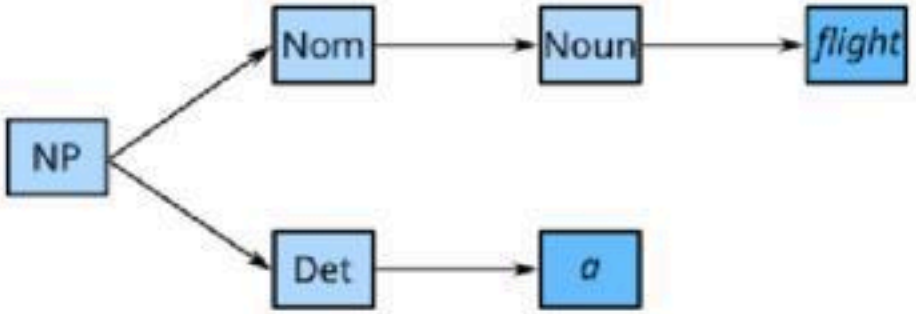
- 类比于语法树的拓展


NP
Nom
Noun
flight
Det
a

终端

CFG中的符号可分为两类：

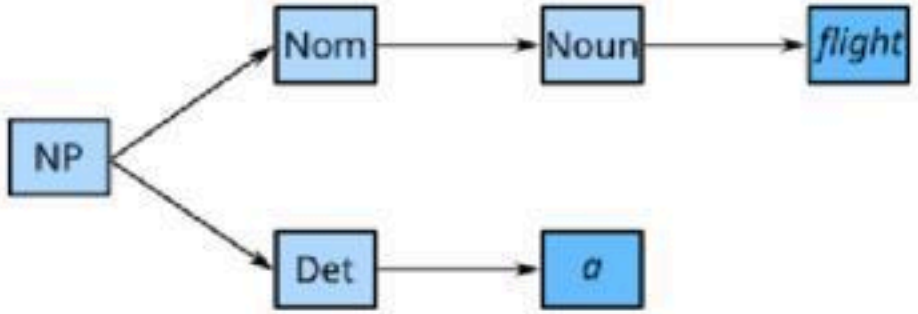
- **终端 terminal**：对应语言（词典）中的字
- **（非终端）节点**：抽象生成出的符号



终端

CFG中的符号可分为两类：

- **终端 terminal**：对应语言（词典）中的字
- **（非终端）节点**：抽象生成出的符号



CFG中的规则可以不断拓展：

- 右侧：可以是单个或多个生成符号、字的有序列表
- 左侧：一定是单个生成符号，表示**聚类或归纳**
  - 指向字的节点表示其**类别或句子成分**

生成、判别

CFG的两种理解：

- **生成 generative**：给定规则，生成句子
- **判别 discriminative**：给定句子，判别结构

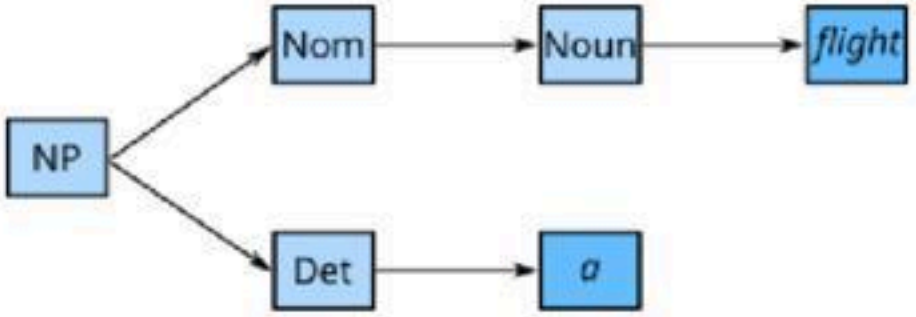
生成、判别

CFG的两种理解：

- **生成 generative**：给定规则，生成句子
- **判别 discriminative**：给定句子，判别结构

生成：“将左边的符号重写成新的符号串”

$$NP : \text{a flight}$$
$$Det\ Nominal$$
$$a\ Noun$$
$$flight$$



解析树

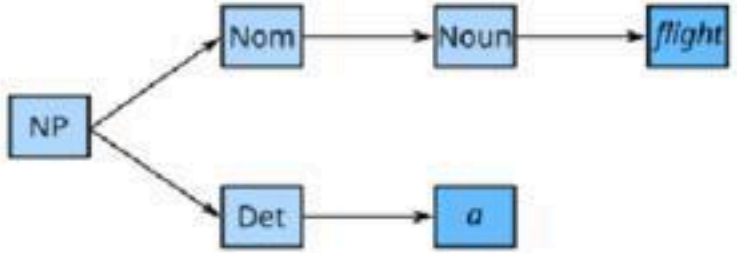
判别句子结构：**解析树 parse tree**的构建

- 句子（的结构解析）与解析树等价

名词短语

```
NP: a flight
Det Nominal
Det Noun
a flight
```

解析树图示



解析树

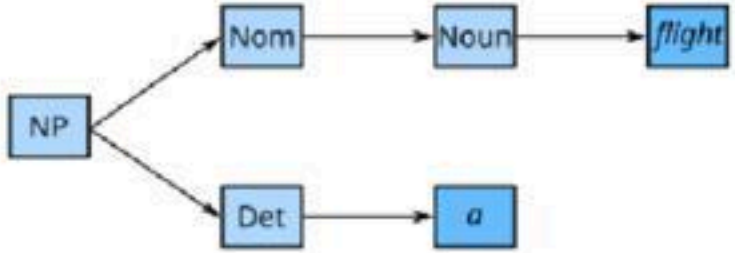
判别句子结构：**解析树 parse tree**的构建

- 句子（的结构解析）与解析树等价

名词短语

```
NP: a flight
Det Nominal
Det Noun
a flight
```

解析树图示



派生 derivation：通过规则扩展符号序列的过程。

- 有向链接可以称为**支配 dominate**
- 树结构有唯一**起始符 start symbol**
  - 常用“S”指代，又可解释为句子本身

动词短语、介词短语

动词短语 verb phrase

$$S \rightarrow NP\ VP \quad \text{I prefer a morning flight}$$

介词短语 prepositional phrase

$$PP \rightarrow Preposition\ NP \quad \text{from Los Angeles}$$

动词短语、介词短语

动词短语 verb phrase

$S \rightarrow NP\ VP$ I prefer a morning flight

$VP \rightarrow Verb\ NP$	prefer a morning flight
$VP \rightarrow Verb\ NP\ PP$	leave Boston in the morning
$VP \rightarrow Verb\ PP$	leaving on Thursday

介词短语 prepositional phrase

$PP \rightarrow Preposition\ NP$ from Los Angeles

ATIS 造句：词典

$$
\begin{aligned}
\text{Noun} &\rightarrow flights|flight|breeze|trip|morning \\
\text{Verb} &\rightarrow is|prefer|like|need|want|fly|do \\
\text{Adjective} &\rightarrow cheapest|non-stop|first|latest|other|direct \\
\text{Pronoun} &\rightarrow me|I|you|it \\
\text{Proper-Noun} &\rightarrow Alaska|Baltimore|Los\,Angeles|Chicago|United|American \\
\text{Determiner} &\rightarrow the|a|an|this|these|that \\
\text{Preposition} &\rightarrow from|to|on|near|in \\
\text{Conjunction} &\rightarrow and|or|but
\end{aligned}
$$

ATIS 造句：例子

语法规则			例子
S	→	NP VP	I + want a morning flight
NP	→	Pronoun	I
	|	Proper-Noun	Los Angeles
	|	Det Nominal	a + flight
Nominal	→	Nominal Noun	morning + flight
	|	Noun	flights

Quiz： ATIS 解析树

如何对“I prefer a morning flight”绘制解析树?

括号表示

括号表示 Bracketed notation：另外一种相对紧凑的表示

```
I prefer a morning flight
```

$[_S[_{NP}[_{Pro}\ \text{I}\ ]][_{VP}[_V\ \text{prefer}\ ][_{NP}[_{Det}\ \text{a}\ ][_{Nom}[_N\ \text{morning}\ ][_{Nom}[_N\ \text{flight}\ ]]]]]]$

括号表示

括号表示 Bracketed notation：另外一种相对紧凑的表示

```
I prefer a morning flight
```

$[_S[_{NP}[_{Pro}\ \text{I}\ ]][_{VP}[_V\ \text{prefer}\ ][_{NP}[_{Det}\ \text{a}\ ][_{Nom}[_N\ \text{morning}\ ][_{Nom}[_N\ \text{flight}\ ]]]]]]$

```
[
    [
        [ I ]
    ]
    [
        [ prefer ]
        [
            [ a ]
            [
```

实验：成分句法分析

HanLP: https://hanlp.hankcs.com/

形式语言

形式语言 formal language：取决于人工定义

- **语法正确的 grammatical**：可以通过形式语言生成的句子

形式语言

形式语言 formal language：取决于人工定义

- **语法正确的 grammatical**：可以通过形式语言生成的句子

生成语法 generative grammar：句子集合完全由形式语言生成

- 例如：CFG
  - 简化问题的讨论
  - 问题：忽略了上下文

CFG 的严格定义 I

N	a set of **non-terminal symbols** (or **variables**)
Σ	a set of **terminal symbols** (disjoint from N)
R	a set of **rules** or productions, each of the form $A \to \beta$
	β is a string of symbols from the infinite set of strings $(\Sigma \cup N)*$
S	a designated **start symbol** and a member of N

CFG 的严格定义 I

N	a set of **non-terminal symbols** (or **variables**)
Σ	a set of **terminal symbols** (disjoint from N)
R	a set of **rules** or productions, each of the form $A \rightarrow \beta$
	β is a string of symbols from the infinite set of strings $(\Sigma \cup N)*$
S	a designated **start symbol** and a member of N

Capital letters like A, B, and S	Non-terminals
S	The start symbol
Lower-case Greek letters like α, β, and γ	Strings drawn from $(\Sigma \cup N)*$
Lower-case Roman letters like u, v, and w	Strings of terminals

CFG 的严格定义 II

直接导出 directly derives

[Hopcroft and Ullman 1979] if $A \rightarrow \beta$ is a production of R and α and γ are any strings in the set $(\Sigma \cup N)*$, then we say that $\alpha A \gamma$ **directly derives** $\alpha\beta\gamma$, or $\alpha A \gamma \Rightarrow \alpha\beta\gamma$.

导出 derives

Let $\alpha 1, \alpha 2, \ldots, \alpha m$ be strings in $(\Sigma \cup N)*, m \geq 1$, such that $\alpha 1 \Rightarrow \alpha 2, \alpha 2 \Rightarrow \alpha 3, \ldots, \alpha m\ 1 \Rightarrow \alpha m$. We say that $\alpha 1$ **derives** αm, or $\alpha 1 \ *\Rightarrow \alpha m$.

CFG 的严格定义 II

直接导出 directly derives

[Hopcroft and Ullman 1979] if A → β is a production of R and α and γ are any strings in the set (Σ∪N)∗, then we say that αAγ **directly derives** αβγ, or αAγ ⇒ αβγ.

导出 derives

Let α1, α2, ..., αm be strings in (Σ∪N)∗,m ≥ 1, such that α1 ⇒ α2,α2 ⇒ α3,...,αm 1 ⇒ αm. We say that α1 **derives** αm, or α1 ∗⇒ αm.

语法G生成的语言

$$\mathcal{L}_G = \{w | w \in \Sigma*, S* \Rightarrow w\}$$

树库

树库

树库 Treebanks：标注好句法的语料库

- 名称来源：句子与解析树等价
- 通用标准：正确性、鲁棒性，专家校正

树库

树库 Treebanks：标注好句法的语料库

- 名称来源：句子与解析树等价
- 通用标准：正确性、鲁棒性，专家校正

常用数库：

- **Penn Treebank**
- **Universal Dependencies**
- **Chinese Treebank**
  - Chinese Treebank 9.0: https://catalog.ldc.upenn.edu/LDC2016T13

Chinese Treebank 9.0

Chinese Treebank 9.0 consists of approximately **two million words** of annotated and parsed text from Chinese newswire, government documents, magazine articles, various broadcast news and broadcast conversation programs, web newsgroups, weblogs, discussion forums, chat messages and transcribed conversational telephone speech.

- There are **3,726 text files** in this release, containing 132,076 sentences, 2,084,387 words, 3,247,331 characters (hanzi or foreign).
- The data is provided in the UTF-8 encoding, and the annotation has Penn Treebank-style **labeled brackets**.
- The data is provided in **four different formats**: raw text, word segmented, POS-tagged, and syntactically bracketed formats.
- All files were automatically verified and **manually checked**.

Chinese Treebank 9.0 样例

文本：青岛优化资本结构促进企业规模扩大

Bracketed

```
<S ID=9901>
( (IP-HLN (NP-PN-SBJ (NR 青岛))
       (VP (VP (VV 优化)
           (NP-OBJ (NN 资本)
               (NN 结构)))
           (VP (VV 促进)
           (NP-OBJ (NP (NN 企业)
                   (NN 规模))
               (NP (NN 扩大))))))))
```

POSTagged

```
<S ID=9901>
青岛_NR 优化_VV 资本_NN 结构_NN 促进_VV 企业_NN 规模_NN 扩大_NN
</S>
```

Review

本章内容

构成句法、语境无关语法。解析树、括号表示。完全切分与词典分词。字典树。评测指标：准确度、精确度、召回率、F度量。

重点：语境无关语法；解析树；词典分词的三种策略；评测指标的计算。

难点：语境无关语法的严格定义；评测指标的原理。

学习目标

- 理解构成句法、语境无关语法的原理。
- 掌握使用CFG对句子构建解析树的方法。
- 掌握解析树对应的括号表示格式。
- 理解完全切分算法与词典分词的三种策略：前向、后向、双向最长匹配。
- 理解字典树的数据结构及应用。
- 理解评测指标的计算方法和原理：准确度、精确度、召回率、F度量。
- 掌握评测词典分词中分隔转换的方法。

问题

如何对“欢迎新老师生前来就餐”绘制解析树？如何转换成对应的括号表示格式？

简述完全切分算法与词典分词的三种策略。

简述字典树的数据结构及应用。

简述评测指标的计算方法和原理：准确度、精确度、召回率、F度量。

简述评测词典分词中分隔转换的方法。

课程项目

(*) 运用生成语法创作短文。

- 中文或英文
- 要求：内容积极、健康，**严禁极端言论**
- 提示：目前可以使用CFG导出句子，之后课程会介绍其他方法