



南京大學  
NANJING UNIVERSITY



# 自然语言处理

## 规则方法

吴震

南京大学人工智能学院  
南京大学自然语言处理研究组

2023年2月

## ● 课程目标

- 了解自然语言处理的发展历程、研究任务/方向
  - 语言模型、文本分类、机器翻译、预训练模型等
- 掌握自然语言处理的基本方法
  - 规则方法、隐马尔科夫模型、神经网络等
- 建立利用自然语言处理知识解决现实问题的能力
  - 自动问答、对话系统等

# 引言：为什么需要规则方法？

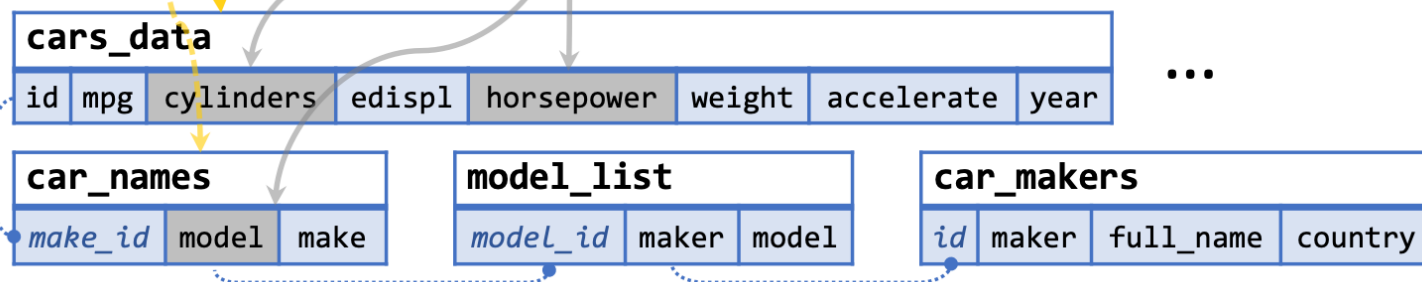
## • Text2SQL

- 输入一句话，将这句话转换成SQL可执行的查询语句

### Natural Language Question:

For the **cars** with 4 cylinders, which **model** has the largest horsepower?

### Schema:



### Desired SQL:

```
SELECT T1.model
FROM car_names AS T1 JOIN cars_data AS T2
ON T1.make_id = T2.id
WHERE T2.cylinders = 4
ORDER BY T2.horsepower DESC LIMIT 1
```

- Question → Column linking (unknown)
- - - - - Question → Table linking (unknown)
- ..... Column → Column foreign keys (known)

- 缺乏足够的标注数据
- 自然语言和SQL语言间复杂的对应关系

机器学习 → 自然语言处理

- 分析型语言
  - 没有或很少有词形变化
  - 词的语法功能由词序和虚词来表达
  - 汉语、藏语等
- 屈折型语言
  - 有词形变化（通过词缀）
  - 词的语法功能由词的形态变化来表达
  - 英语、德语、法语等

- 黏着型语言
  - 有词形变化和附加成分
  - 词的语法功能由词的形态变化和附加成分来表达
  - 日语、芬兰语等

以日语"食べる"为例:

食べる Tabe ru - "吃" (基本形、将来时)

食べさせる Tabe sase ru - "吃" + 使役助动词 - 使/要求(某人)吃

食べさせられる Tabe sase rare ru - "吃" + 使役助动词 + 被动助动词 - 被(其他人)要求(我)吃

食べさせられない Tabe sase rare nai - "吃" + 使役助动词 + 被动助动词 + 否定助动词 - 不被(其他人)要求(我)吃

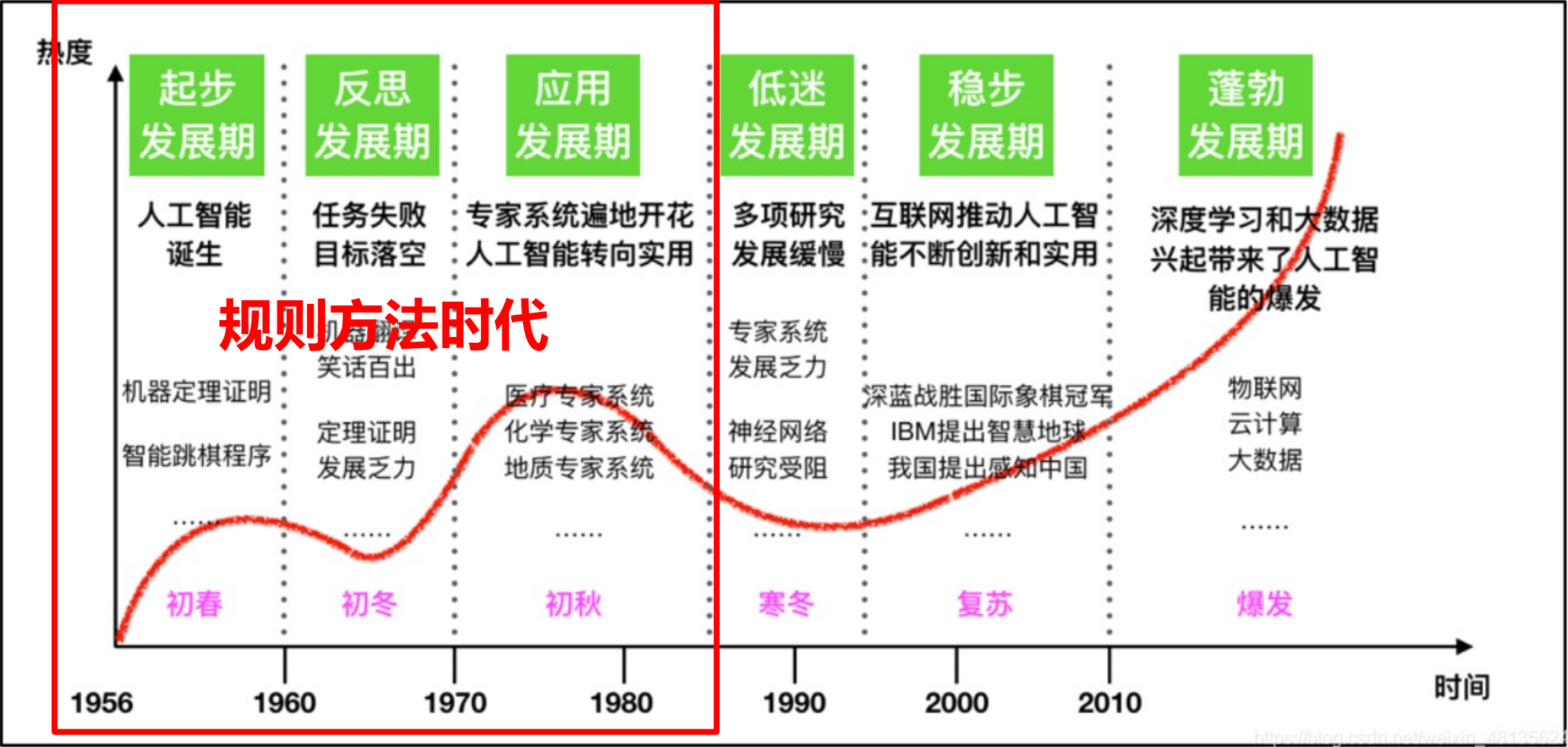
食べさせられなかった Tabe sase rare na katta - "吃" + 使役助动词 + 被动助动词 + 否定助动词 + 过去助动词 - 曾不被(其他人)要求(我)吃

# 自然语言分类-基于“主动宾”位置

- SVO型（主-动-宾）
  - 英语、汉语等
- SOV型（主-宾-动）
  - 日语等
- VSO型（动-主-宾）
  - 古阿拉伯语等

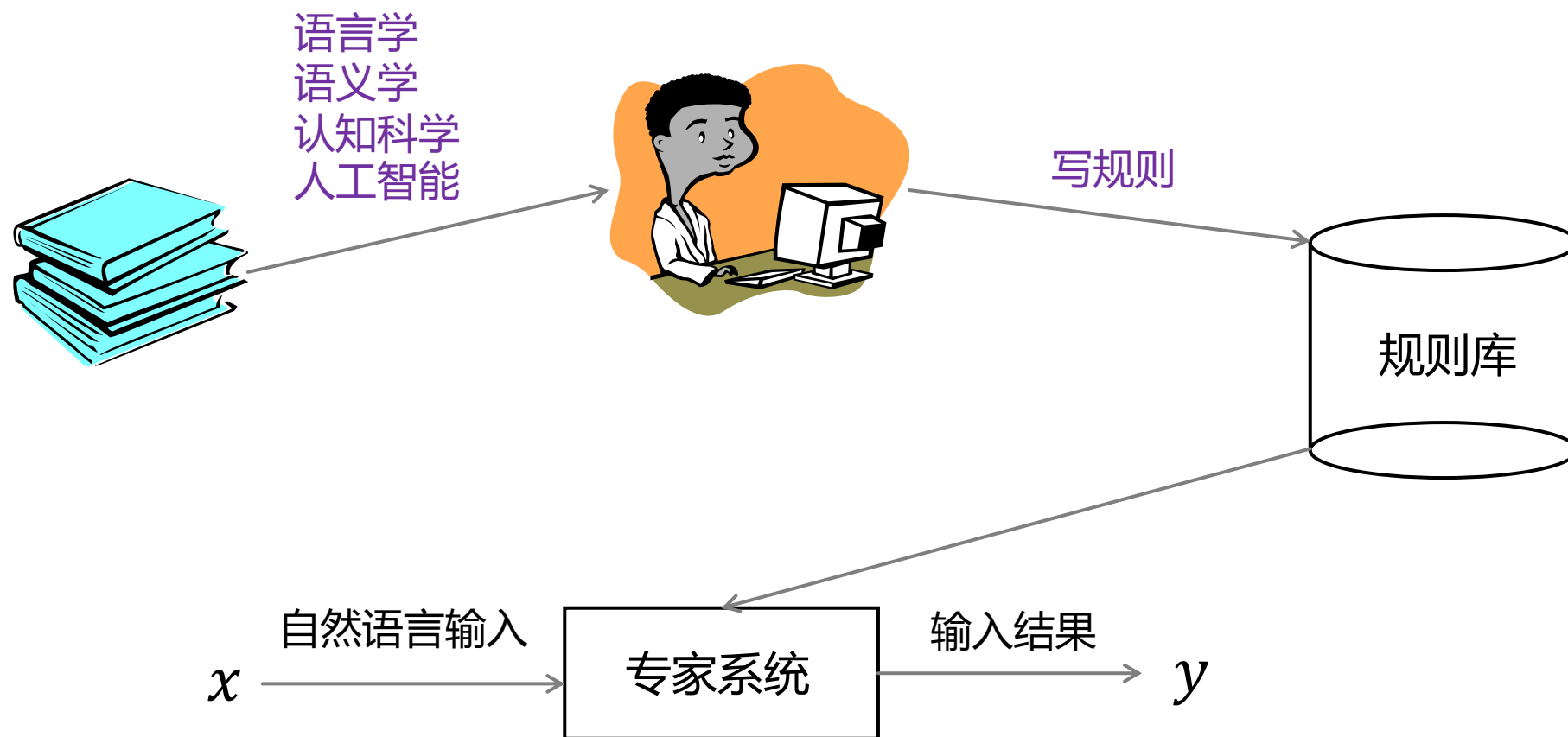
对翻译的影响？





- 以**规则形式**表示语言知识
- 基于规则的知识表示和推理
- 强调人对语言知识的理性整理（知识工程）
- 受计算语言学理论指导
- 语言规则（数据）与程序分离，程序体现为规则语言的解释器！





- 词法分析
- 句法分析
- 语义分析
- .....

- 形态还原（针对英语、德语、法语等）
  - 把句子中的词还原成它们的基本词形（原形）
- 分词（针对汉语、日语等）
  - 识别出句子中的词
- 词性标注
  - 为句子中的词标上预定义类别集合中的类
- 命名实体识别
  - 识别出句子中的人名、地名、机构名等

- 把句子中的词还原成原形，作为词的其它信息的索引（词典、个性规则）
- 构词特点
  - 屈折变化：词尾和词形变化，词性不变。如：
    - study, studied, studied, studying
    - speak, spoke, spoken, speaking
  - 派生变化：加前缀和后缀，词性发生变化。如：
    - friend, friendly, friendship,...
  - 复合变化：多个单词以某种方式组合成一个词。
- 还原规则
  - 通用规则：变化有规律
  - 个性规则：变化无规律

## ● 规则动词还原举例

- \*s -> \* (SINGULAR3)
- \*es -> \* (SINGULAR3)
- \*ies -> \*y (SINGULAR3)
- \*ing -> \* (VING)
- \*ing -> \*e (VING)
- \*ying -> \*ie (VING)
- \*??ing -> \*? (VING)
- \*ed -> \* (PAST)(VEN)
- \*ed -> \*e (PAST)(VEN)
- \*ied -> \*y (PAST)(VEN)
- \*??ed -> \*? (PAST)(VEN)

- 不规则动词还原举例
  - went -> go (PAST)
  - gone -> go (VEN)
  - sat -> sit (PAST) (VEN)



- 基本算法（给定一段文本）
  1. 输入一个单词
  2. 如果原型词典里有该词，输出该词及其属性
  3. 如果有该词的还原规则，并且词典里有还原后的词，则输出还原后的词及其属性；否则，调用<未登录词模块>
  4. 如果输入中还有单词，转1；否则，结束。

- 词是语言中最小的能独立运用的单位，也是语言信息处理的基本单位。
- 分词是指根据某个分词规范，把一个“字”串划分成“词”串。
- 分词规范
  - 由于单字词的存在，有时无法区分：
    - 词与语素：猪肉（鸭肉）、猪在奔跑、肉很香
    - 词与词组：黑布、黑板
  - 信息处理用现代汉语分词规范：GB-13715（1992）
  - 具体应用系统可根据各自的需求制定规范

- 交集型歧义

- ABC切分成AB/C或A/BC
- 如：“**和**平等”
  - “独立/自主/**和/平等**/独立/的/原则”
  - “讨论/战争/与/**和平/等**/问题”

- 组合型歧义

- AB切分成AB或A/B
- 如：“**马上**”
  - “他/骑/在/**马/上**”
  - “**马上**/过来”

- 混合型歧义
  - 由交集型歧义和组合型歧义嵌套与交叉而成
  - 如：“**得到达**”（交集型、组合型）
    - “我/今晚/**得/到达**/南京”
    - “我/**得到/达**克宁/了 ”
    - “我/**得/到/达**克宁/公司/去”

- 伪歧义与真歧义

- 伪歧义字段指在任何情况下只有一种切分

- “**为人民**” 只有一种切分: “**为/人民**” , 如: “为/人民/服务”
- 根据歧义字段本身就能消歧

- 真歧义字段指在不同的情况下有多种切分

- “**从小学**” 可以有多种切分：
  - “**从小/学**” , 如: “从小/学/电脑” ( “从小” 是切分成 “从小” 还是 “从/小” 要根据分词规范! )
  - “**从/小学**” , 如: “他/从/小学/毕业/后”
- 根据歧义字段的上下文来消歧

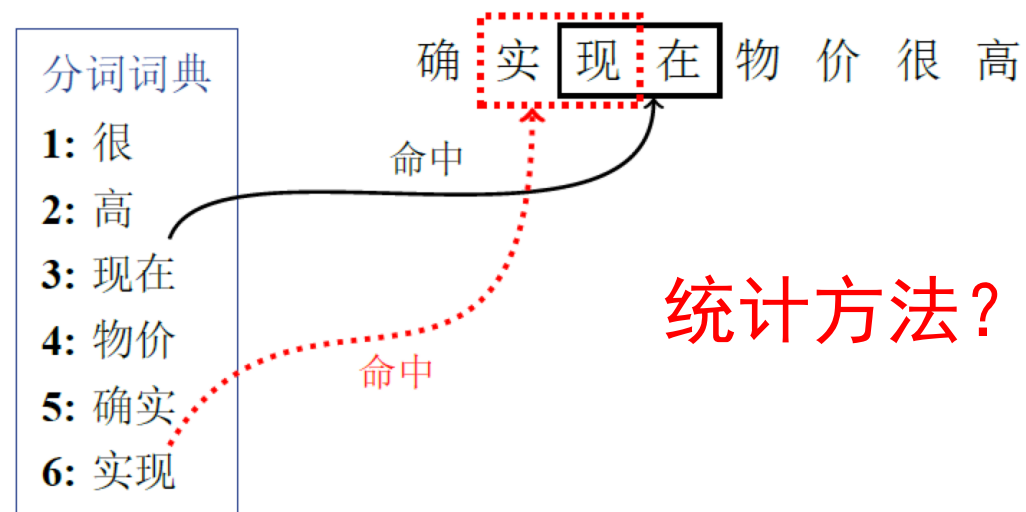
- 正向最大匹配(FMM)或逆向最大匹配(RMM)
  - 从左至右(FMM)或从右至左(RMM)，取最长的词
- 双向最大匹配
  - 分别采用FMM和RMM进行分词，能发现交集型歧义（“幼儿园/地/节目”和“幼儿/园地/节目”）
  - 如果结果一致，则认为成功；否则，采用消歧规则进行消歧
- 正向最大、逆向最小匹配
  - 正向采用FMM，逆向采用最短词，能发现组合型歧义（“他/骑/在/马上”和“他/骑/在/马/上”）
- 逐词遍历匹配
  - 在全句中取最长的词，去掉之，对剩下字符串重复该过程
- 设立切分标记
  - 收集词首字和词尾字，先把句子分成较小单位，再用某些方法切分
- 全切分
  - 获得所有可能的切分，选择最可能的切分



# 基于规则的歧义字段消歧方法

- 利用歧义字串、前驱字串和后继字串的句法、语义和语用信息
  - 句法信息
    - “阵风”：根据前面是否有数词来消歧。“一/阵/风/吹/过/来”、“今天/有/阵风”
  - 语义信息
    - “了解”：“他/学会/了/解/数学/难题”（“难题”一般是“解”而不是“了解”）
  - 语用信息
    - “拍卖”：“乒乓球拍卖完了”，要根据场景（上下文）来确定
- 规则的粒度
  - 基于具体的词（个性规则）
  - 基于词类、词义类（共性规则）

# 基于规则的歧义字段消歧方法



统计方法?

交集型分词歧义

# 分词带来的问题

---

- 组成词的字的信息丢失
- 错误的分词影响后续的工作
- 不同分词规范的分词造成分词结果不一致

不分词行吗？

- 定义：为句子中的词标上预定义类别集合中的类（词性）
- 目标：为后续的句法/语义分析提供必要的信息
- 标注体系
- 标注方法



- |       |      |        |
|-------|------|--------|
| 1 名词  | 5 代词 | 9 量词   |
| 2 动词  | 6 介词 | 10 助词  |
| 3 形容词 | 7 连词 | 11 感叹词 |
| 4 副词  | 8 数词 | 12 拟声词 |

一个以义为纲的词汇分类体系  
——《现代汉语分类词典》\*

- 词的分类
  - 按形态和句法功能（句法相关性）
  - 按表达的意思（语义相关性）
  - 兼顾上述二者

- 开放类 ( open class , 每类词数不限 )
  - Nouns
    - 句法上：可作物主、可有限定词、有复数形式
    - 语义上：人名、地名和物名等
  - Verbs
    - 句法上：作谓语、有几种词形变化
    - 语义上：动作、过程 ( 一系列动作 )
  - Adjectives
    - 句法上：修饰Nouns等
    - 语义上：性质
  - Adverbs
    - 句法上：修饰Verbs等
    - 语义上：方向、程度、方式、时间



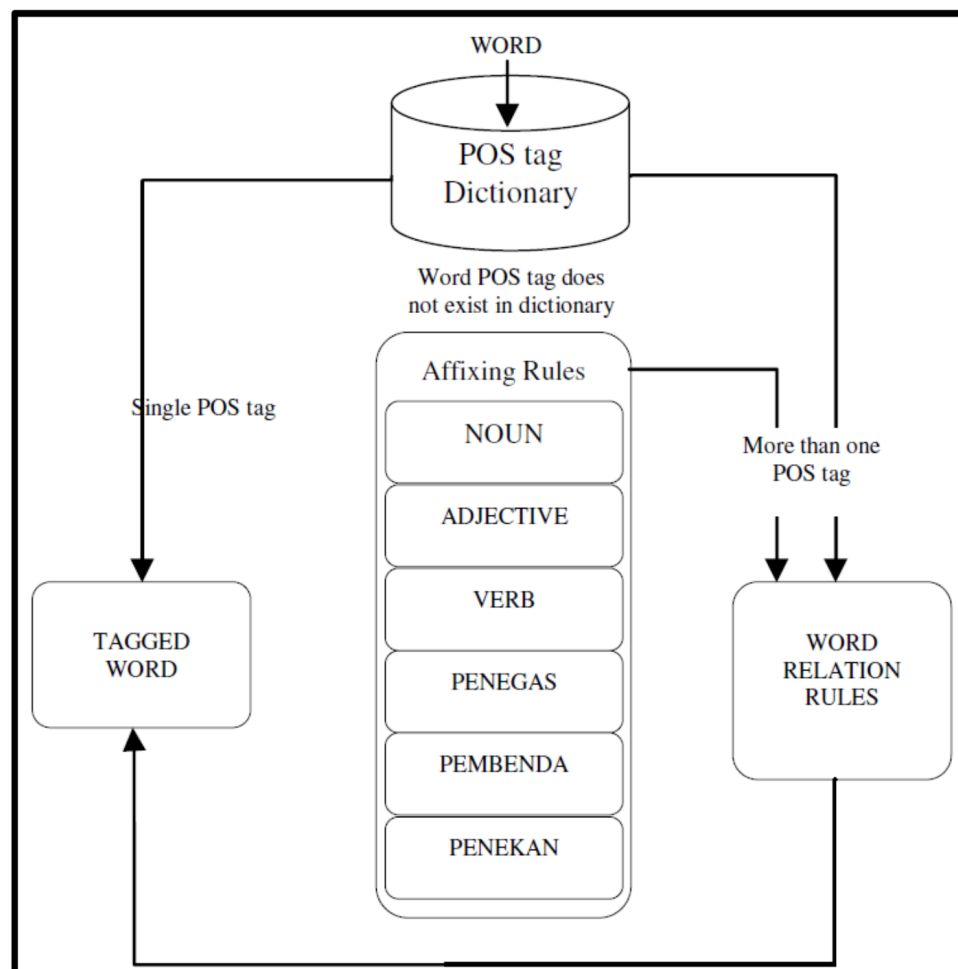
- 封闭类 ( closed class , function words , 每类词数有限 )
  - Determiners
  - Pronouns
  - Prepositions
  - Conjunctions
  - Auxiliary verbs
  - Particles ( if、 not、 ... )
  - Numerals

- 兼类词

- 一个词具有两个或者两个以上的词性
- 英文的Brown语料库中，10.4%的词是兼类词。例如：
  - The **back** door
  - On my **back**
  - Promise to **back** the bill
- 汉语兼类词，例如：
  - 把门**锁**上      买了一把**锁**
  - 他**研究**xx      他的**研究**工作...
  - 由于缺少词形变化，汉语的兼类词更多！

- 规则方法

- 词典和规则提供候选词性
- 消歧规则进行消歧

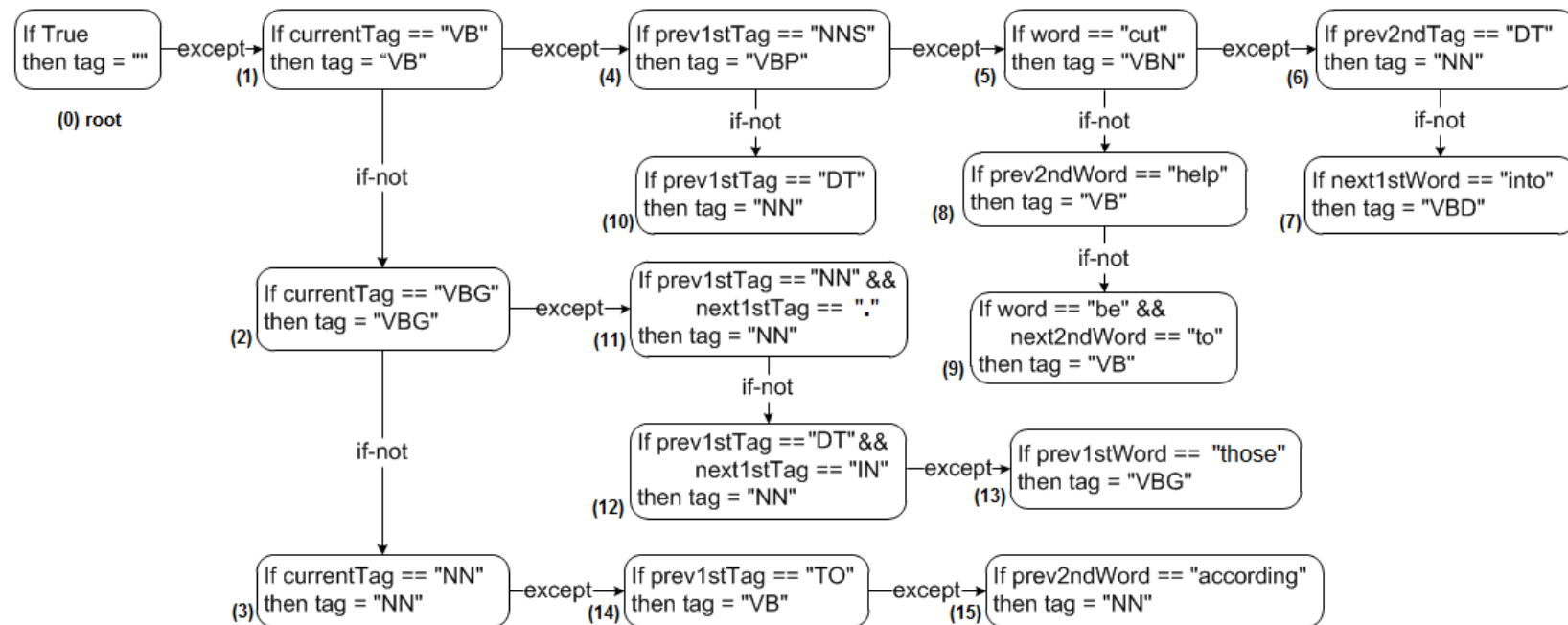


## ● 统计方法

- 选择最可能的词性结果
- 训练用语料库（已标注）

## ● 基于转换的错误驱动方法

- 大规模标注语料库
- 统计学习规则
- 用规则方法进行标注



RDRPOSTagger：以二叉树的形式自动构建标注规则  
<https://github.com/datquocnguyen/RDRPOSTagger>

# 命名实体识别 (NER)

- 定义：识别出一句话中的所有命名实体
  - 找到实体的边界
  - 确定实体类型
- 目的：识别出实体，为后续的语义分析提供支持

人名          机构名          地名  
└──┬──┘      └──┬──┘      └──┬──┘  
小明   在   南京大学   的   方肇周体育馆   看了一场比赛

- 3大类
  - 实体类、时间类、数字类
- 7小类
  - 人名、地名、机构名、时间、日期、货币量、百分数
- 其他体系（根据需要）

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	<b>Turing</b> is a giant of computer science.
Organization	ORG	companies, sports teams	The <b>IPCC</b> warned about the cyclone.
Location	LOC	regions, mountains, seas	<b>Mt. Sanitas</b> is in <b>Sunshine Canyon</b> .
Geo-Political Entity	GPE	countries, states	<b>Palo Alto</b> is raising the fees for parking.





# 基于规则的命名实体识别

- 核心思想
  - 匹配
  - 依赖词典、模板、正则表达式
- 人名限制成分
  - 身份词：深度学习专家Hinton
  - 指界词：Hinton参加/同意/反对
  - 标点符号：2018年图灵奖得主是Hinton、Bengio、Yann LeCun。

- 基于实体词表的匹配识别
  - 专家总结实体词表，利用词表进行匹配
  - 优点：速度快
  - 缺点：覆盖率有限，人力总结
  - 适合垂直领域，如医疗、金融、法律等

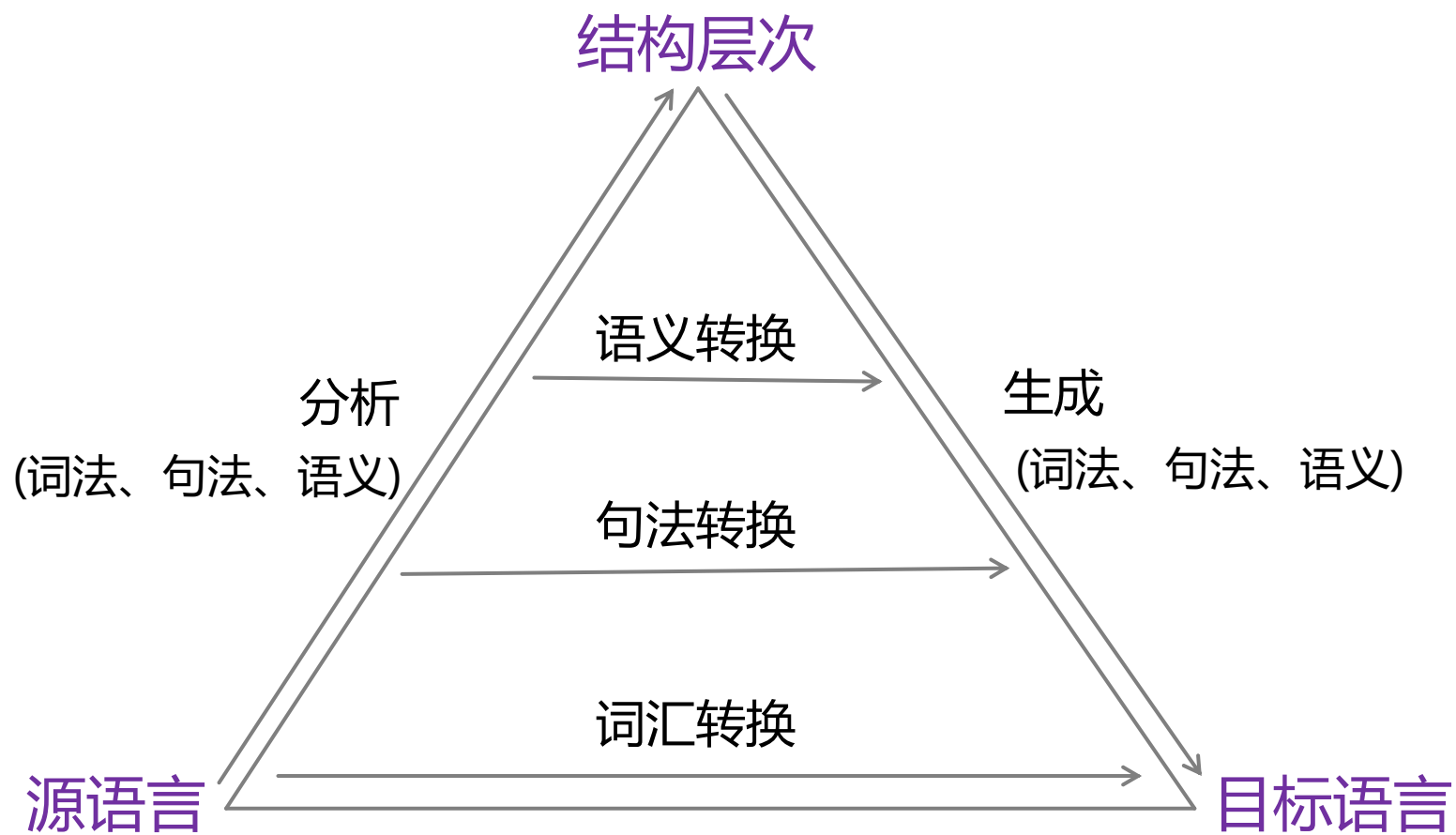
变化的实体（邮箱）怎么识别？？？

- 基于规则模板的匹配识别
  - 可以作为实体词表识别的补充
  - 分析实体词或者属性值的构词规则，并构建规则模板（正则表达式）
    - Email的表现形式通常为xxxx@xxx.com，利用 “`^\w+([-+.] \w+ )@\w+([-.] \w+ ).\w+([-.] \w+ )*$`” 来匹配Email地址
    - 利用 “`\d{4}[年-]\d{1,2}[月-]\d{1,2}日`” 的正则模板表达式来提取日期

- 机器翻译 ( Machine Translation ) 是一个将源语言的句子  $x$  翻译成目标语言句子  $y$  ( 译文 ) 的任务。



- 分析
  - 将源语言句子解析成一种深层的结构表示
- 转换
  - 将源语言句子的深层结构表示转换成目标语言的深层结构表示
- 生成
  - 根据目标语言的深层结构表示生成对应的目标语言句子



- 基于词的转换翻译
- 基于句法结构转换的翻译
- 基于语义转换的翻译
- 基于中间语言（Interlingua）的翻译

- 翻译过程
  - 词法分析（源语言）
  - 译词选择
  - 词序调整
  - 形态（词形变化）生成
- 翻译所需的知识
  - 词法规则（源语言）
  - 双语词典及规则
  - 调序规则
  - 形态生成规则



- 词汇转换规则

- 他 → He , 在 → in , 北京 → Beijing , 工作 → works

源语言

他

在

北京

工作

目标语言



He



in



Beijing



works



- 翻译过程
  - 句法分析（源语言）
  - 递归地利用一组“树-树”的转换规则，把源语言的句法树转换成目标语言的句法树
  - 从目标语言的句法树生成目标语言句子。

- 分析规则

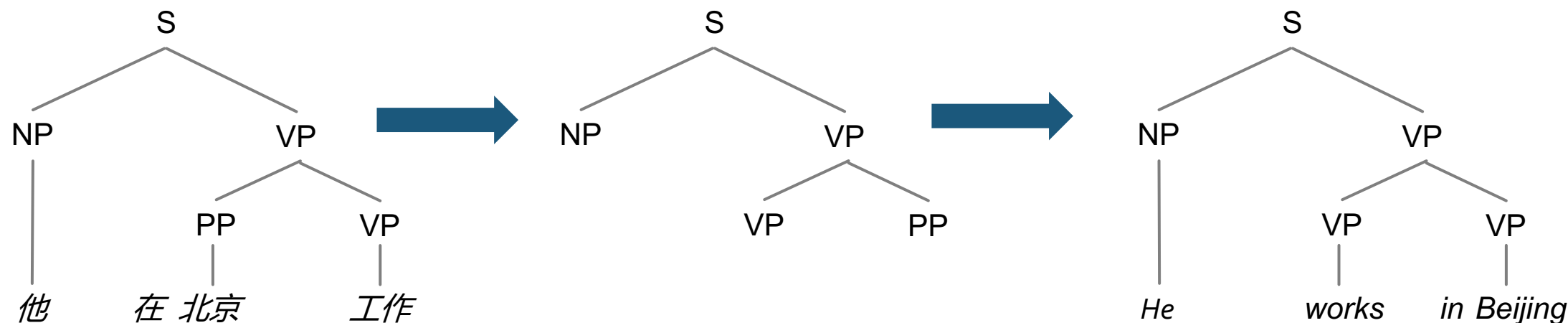
- NP → 他, PP → 在北京, VP → 工作, VP → PP VP, S → NP VP

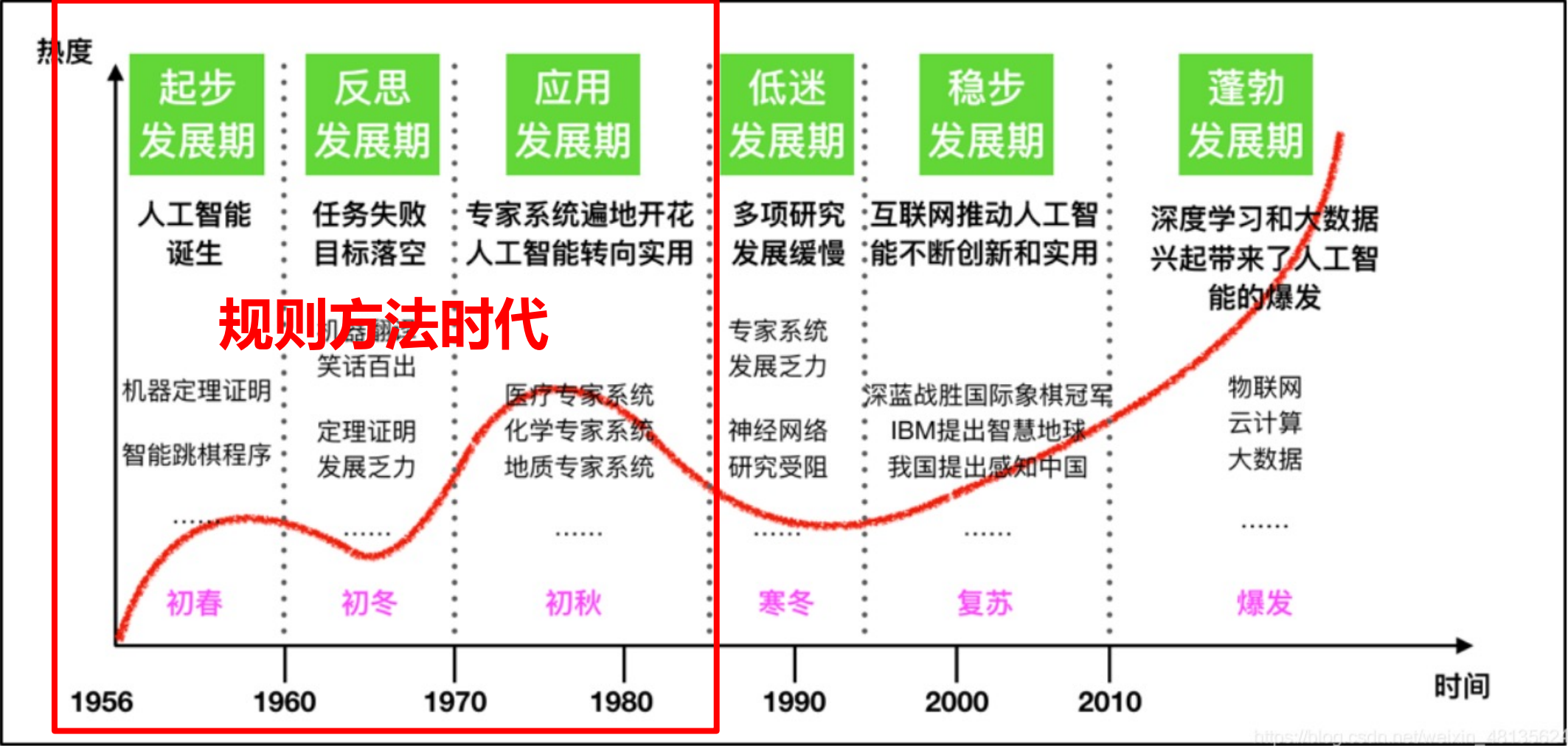
- 句法转换规则

- S (NP VP) → S (NP VP), VP (PP VP) → VP(VP PP)

- 词汇转换规则

- 他 → He, 在 → in, 北京 → Beijing, 工作 → works





- 规则质量依赖于语言学家的知识和经验，获取成本高
- 规则之间容易发生冲突
- 大规模规则系统维护难度大



"Anytime a linguist leaves the group the recognition rate goes up."

- Frederick Jelinek, 1988

Frederick Jelinek (1932-2010)

Researcher in Information Theory, Speech Recognition, and Natural Language Processing

Professor at Cornell 1962-1974

Head of Continuous Speech Recognition group, IBM T.J. Watson 1972-1993

Head of Center for Language and Speech Processing, JHU 1993-2010

Thank you !  
Q&A

