自然语言处理

Natural Language Processing(NLP)

陈家骏，戴新宇

chenjj@nlp.nju.edu.cn

dxy@nlp.nju.edu.cn

http://nlp.nju.edu.cn

主要内容（1）

- 自然语言处理概述
  - 什么是自然语言处理
  - 自然语言处理技术的应用
  - 自然语言处理的基本策略和实现方法
  - 自然语言处理的难点
  - 自然语言处理所涉及的学科

主要内容（2）

- 基于规则的自然语言处理方法（**理性方法**，传统方法）
  - 基于词典和规则的分词（汉语、日语）
  - 基于CFG（上下文无关文法）的句法表示及其分析技术
  - 基于扩充的CFG（复杂特征集、合一运算）的句法表示及其分析技术
  - 词义及句义表示：基于逻辑形式和格语法的句义分析
  - 基于规则的机器翻译

主要内容（3）

- 基于语料库和统计学习的自然语言处理方法（**经验方法**）
  - 语言模型（N元文法）
  - 分词、词性标注（序列化标注模型）
  - 句法分析（概率上下文无关模型）
  - 文本分类（朴素贝叶斯模型、最大熵模型）
  - 机器翻译（IBM Model）

所需的前导知识

- 编译技术
- 概率与统计

参考书籍

- 刘群等译，*自然语言理解（第二版）*，电子工业出版社，2005
- 苑春法等译，*统计自然语言处理基础*，电子工业出版社，2005
- 冯志伟等译，*自然语言处理综论*，电子工业出版社，2005
- 黄昌宁等，*语料库语言学*，商务印书馆，2002
- 冯志伟，*计算语言学基础*，商务印书馆，2001
- 余士文，*计算语言学概论*，商务印书馆，2003
- 姚天顺，*自然语言理解——一种让机器懂得人类语言的研究（第2版）*，清华大学出版社，2002
- 宗成庆，*统计自然语言处理*，清华大学出版社，2008
- 王小捷等，*自然语言处理技术基础*，北京邮电大学出版社,2002
- 刘颖，*计算语言学*，清华大学出版社，2002

- Bonnie J. Dorr, et al, ***Survey of Current Paradigms in Machine Translation***， Technical Report LAMP-TR-027, Language and Media Processing Lab, University of Maryland.
- Hutchins WJ, ***Machine Translation: Past, Present, Future***. Chichester: Ellis Horwood, 1986
- Arturo Trujillo, Translation Engines: ***Techniques for Machine Translation***, Springer-Verlag London Limited 1999
- Peter F. Brown, et al., ***A Statistical Approach to MT***, Computational Linguistics, 1990,16(2)
- P.F. Brown, et al., ***The Mathematics of Statistical Machine Translation: Parameter Estimation***, Computational Linguistics, 1993, 19(2)
- 赵铁军等，***机器翻译原理***，哈尔滨工业大学出版社，2000
- 宗成庆等译，***统计机器翻译***，电子工业出版社，2012

课程考核

- Projects
- 提交要求（每个project）
  - 报告（说明基本做法）
  - 源程序及可运行的程序

自然语言处理概述

什么是自然语言处理

- 充分利用信息将会给人们带来巨大的收益，而大量的信息以**自然语言**形式（英语、汉语等）存在。
- **如何有效地获取和利用以自然语言形式出现的信息？**
- 自然语言处理（Natural Language Processing，简称NLP）是指用计算机对语言信息进行处理的方法和技术。
- 与NLP相近的两个研究领域：
  - 自然语言理解(NLU)：强调对语言含义和意图的深层次解释
  - 计算语言学(CL)：强调可计算的语言理论

NLP技术的应用

- 机器翻译
- 自动摘要
- 文本分类
- 信息检索
- 信息抽取
- 自动问答
- 情感分析
- ……

机器翻译（Machine Translation）

- 机器翻译（Machine Translation，简称MT）是指利用计算机实现自然语言（英语、汉语等）之间的自动翻译。
  - 文本机器翻译
  - 语音机器翻译
- 机器辅助翻译（Machine Aided Translation或Computer Aided Translation，简称MAT或CAT）
  - 翻译记忆体（Translation Memory，简称TM）
  - 双语对照的文本编辑
  - ...

自动摘要（Text Summarization）

- 利用计算机自动地从原始文档中提取全面准确地反映该文档中心内容的简单连贯的短文。
- 压缩比

文本分类（Text Classification）

- 利用计算机将一篇文章归于预先给定的某一类或某几类的过程。
- 可用于信息过滤（Information Filtering）

信息检索（Information Retrieval，IR）

- 主题相关的文本获取。
- google、百度、...（基于关键词的）

信息抽取（Information Extraction，IE）

- 主题相关的信息获取
- 信息抽取是指从非结构化或半结构化的自然语言文本中提取出与某个主题相关的结构化信息。
- IE对数据挖掘的支持

信息抽取实例:会议报道(人民日报1998-03-09)

新华社北京３月８日电（记者李术峰）：中国农工民主党第十二届中央常务委员会第一次会议今天在北京召开。

会议研究通过了贯彻落实“两会”精神的有关决定，审议通过了中国农工民主党中央１９９８年工作要点（草案），并任命了中央副秘书长。

农工民主党中央主席蒋正华主持了会议，他说，农工民主党有１００多名党员作为代表和委员参加了今年的“两会”，各位党员要认真履行代表和委员的职责，开好会，在１９９８年的工作中认真贯彻“两会”精神，加强农工民主党的自身建设，推动事业进一步发展，为建设有中国特色社会主义事业作出新的贡献。

会前，农工民主党中央邀请参加“两会”的来自全国各省、自治区、直辖市的农工民主党党员进行了联谊活动。

信息抽取的结果

会 议 时 间 Time	１９９８年3月8日	
会 议 地 点 Spot	北京	
会议召集者/ 主 持 人Convener	个 人 姓 名 / 团 体名称 Name	蒋正华
	机 构 、 职 位 Org/Post	主席，农工民主党中央
会 议 名 / 标 题Conf-Title	中国农工民主党第十二届中央常务委员会第一次会议	

自动问答（Question Answering，QA）

- 针对用户提出的问题，给出具体的答案。
- Apple的Siri、IBM的Watson机器人、百度的“知道”、…

情感分析（Sentiment Analysis或 Opinion Analysis ）

- 分析文章对某个对象的态度是正面还是负面。
  - 公共关系：舆情分析
  - 市场决策：产品意见调查
  - ……

自然语言处理的主要任务

- 语言分析：分析语言表达的结构和含义
  - 词法分析：形态还原、词性标注、命名实体识别、分词（汉语）等
  - 句法分析：组块分析、结构分析、依存分析
  - 语义分析：词义、句义（逻辑、格关系、...）、篇章（上下文分）（指代、实体关系）
- 语言生成：从内部表示生成语言表达
- 多语言处理：语言之间的对齐、转换
- 不同的应用对上述任务有不同的要求。
  - 机器翻译需要NLP各方面的方法和技术支持，是NLP的典型应用，它几乎涵盖了NLP各个任务。

自然语言处理的实现方法

- 基于语言规则的理性方法（Rationalist approach）
  - 基于以规则形式表达的语言知识（词、句法、语义以及转换、生成）进行推理。
  - 强调人对语言知识的理性整理。
  - Chomsky：先天语言能力，主宰1960－1985
- 基于语料库和统计学习的经验方法（Empiricist approach）
  - 以大规模语料库（单语和双语）为语言知识基础。
  - 利用统计学习方法自动获取和运用隐含在语料库中的知识
  - 知识体现为一系列统计数据（参数）

- 混合方法
  - 理性方法的优、缺点
    - 相应的语言学理论基础好
    - 描述精确
    - 效率高
    - 知识获取困难（高级劳动）
    - 鲁棒性（适应性）差：不完备的规则系统将导致推理的失败
    - 知识扩充困难，很难保证规则之间的一致性
  - 经验方法的优、缺点
    - 知识获取容易（低级劳动）
    - 鲁棒性好：概率大的作为结果
    - 扩充容易、一致性容易维护
    - 相应的语言学理论基础差
    - 缺乏对语言学知识的深入利用，过于机械
    - 效率低
  - 利用各家之长，相互融合

自然语言处理的难点

- 歧义处理
  - 有限的词汇和规则表达复杂的、无限的语言
- 语言知识的表示、获取和运用
- 成语和惯用型的处理
- 对语言的灵活性和动态性的处理
  - 灵活性：同一个意图的不同表达，甚至包含错误的语法等
  - 动态性：语言在不断的变化，如：新词等
- 上下文和世界知识（语言无关）的利用和处理

汉语处理的难点

- 缺乏计算语言学的句法/语义理论，大都借用基于西方语言的句法/语义理论
- 词法分析
  - 分词
  - 词性标注难
- 句法分析
  - 主动词识别难
  - 词法分类与句法功能对应差
- 语义分析
  - 句法结构与句义对应差
  - 时体态确定难 （汉语无形态变化）
- 资源（语料库）缺乏

自然语言处理所涉及的学科

- 计算语言学：各种语法、语义理论
- 计算机科学（包括人工智能）
- 数学：逻辑、概率与统计、信息论，等
- 哲学（认知学）
- 心理学
- ……

基于规则的自然语言处理方法

（ 理性方法，传统方法）

概述

- 强调对语言知识的理性整理（知识工程）
- 受计算语言学理论指导
- 基于规则的知识表示和推导
- 语言处理规则（数据）与程序分离，程序体现为规则语言的解释器！

自然语言的分类（基于形态结构）

- 分析型语言
  - 词形变化很少
  - 没有表示词的语法功能的附加成分，由词序和虚词表示词之间的语法关系
  - 汉语、藏语等
- 黏着型语言
  - 有词形变化
  - 词的语法意义（功能）由附加成分表达
  - 芬兰语、日语等
- 屈折型语言
  - 有词形变化
  - 词的语法意义由词的形态变化来表示
  - 英语、德语、法语等
- 另外，还可以按SVO型（主－动－宾）、VSO型（动－主－宾）和SOV 型(主－宾－动) 分类

词法分析

- 形态还原（针对英语、德语、法语等）
  - 把句子中的词还原成基本词形，作为词的其它信息（词典、个性规则）的索引。
- 词性标注
  - 为句子中的词标上预定义类别集合（标注集）中的类。
- 分词（针对汉语、日语等）
  - 识别出句子中的词。
- 命名实体识别
  - 人名
  - 地名
  - 机构名

形态还原（英语）

- 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

形态还原规则举例

- 英语“规则动词”还原
  - *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)

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

形态还原算法

1. 输入一个单词
2. 如果词典里有该词，输出该词及其属性，转4，否则，转3
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，转4，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。

Proj. 1 实现一个英语单词还原工具。

（词典：http://nlp.nju.edu.cn/MT_Lecture/dic_ec.rar）

词性标注

- 为句子中的词标上预定义类别集合（标注集）中的类，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

词性标注体系

- 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
- 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*...，　*研究*工作
  - 汉语词的兼类更多？与所采用的分类体系是否有关？
- 为什么要分类？分类带来的问题？

英语词的分类

- 开放类（open class）
  - Nouns
    - 句法上：可有限定词、可作物主、有复数形式
    - 语义上：人名、地名和物名
  - Verbs
    - 句法上：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）
  - Determiners
  - Pronouns
  - Prepositions
  - Conjunctions
  - Auxiliary verbs
  - Particles（if、not、...）
  - Numerals

词性标注方法

- 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
  - 选择最可能的标注
  - 训练用语料库（已标注）
- 基于转换学习的方法
  - 统计学习规则
  - 用规则方法进行标注

汉语分词（切分）

- 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”串分成“词”串。
- 分词规范
  - 难以确定何谓汉语的“词”
    - 单字词与语素的界定：猪肉、牛肉
    - 词与短语（词组）的界定：黑板、黑布
  - 信息处理用现代汉语分词规范：GB-13715（1992）
  - 具体系统可根据各自的需求制定规范

切分歧义及歧义字段的种类

- 交集型歧义字段
  - 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”
- 组合型歧义字段
  - 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：“太平淡”（组合型、交集型）
    - “这/墙/抹/得/太/平/了”（组合型）
    - “即使/太平/时期/也/不/应该/放松/警惕”（组合型）
    - “这/篇/文章/写/得/太/平淡/了”（交集型）

南京市长江大桥…

南京市长江二桥…

- 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

分词方法

一般通过分词词典和分词规则库进行分词。主要方法有：

- 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
  - 会忽略“词中有词”的现象：“幼儿园 地 节目”
- 双向最大匹配
  - 分别采用FMM和RMM进行分词
  - 如果结果一致，则认为成功；否则，
  - 采用消歧规则进行消歧（交集型歧义）：
- 正向最大、逆向最小匹配
  - 发现组合型歧义
- 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，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，选择最大可能的切分

基于规则的歧义字段消歧方法

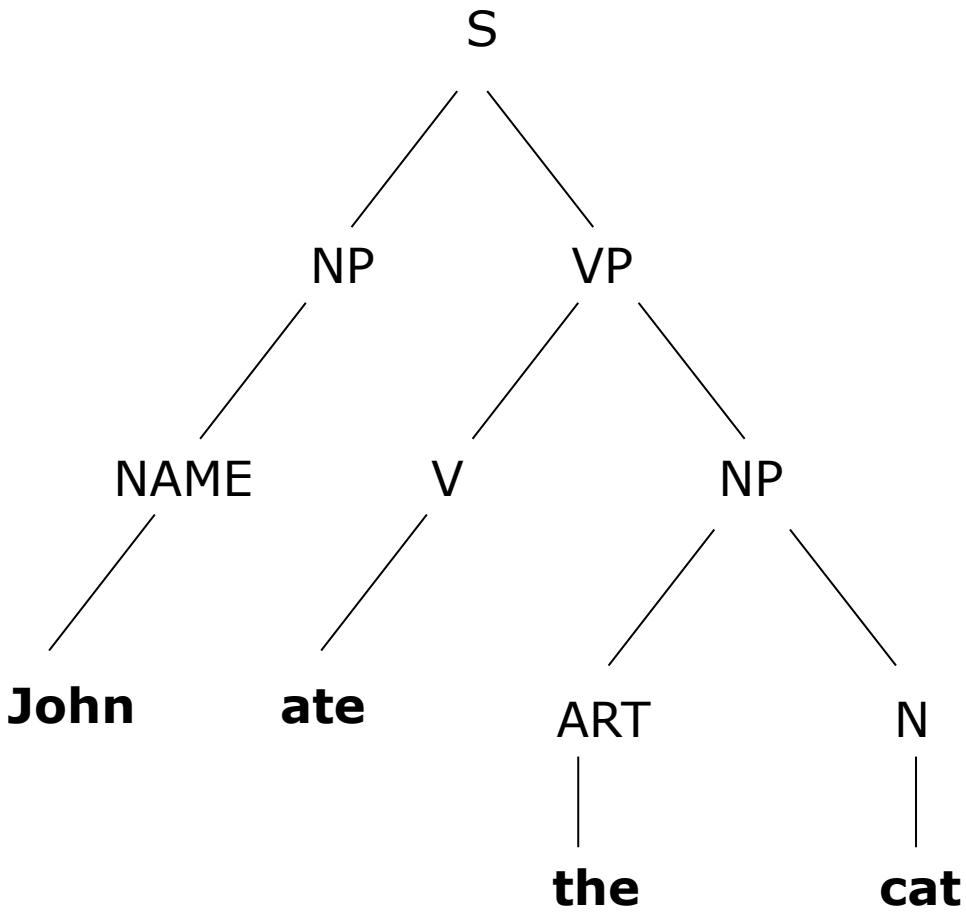
- 利用歧义字串、前驱字串和后继字串的句法、语义和语用信息：
  - 句法信息
    - “阵风”：根据前面是否有数词来消歧。“一/阵/风/吹/过/来”、“今天/有/阵风”
  - 语义信息
    - “了解”：“他/学会/了/解/数学/难题”（“难题”一般是“解”而不是“了解”，另外，还有“学会”）
  - 语用信息
    - “拍卖”：“乒乓球拍卖完了”，要根据场景（上下文）来确定
- 规则的粒度
  - 基于词（个性规则）
  - 基于词类、词义（共性规则）

Proj. 2 实现一个基于词典与规则的汉语自动分词系统。
(词典：http://nlp.nju.edu.cn/MT_Lecture/dic_ce.rar)

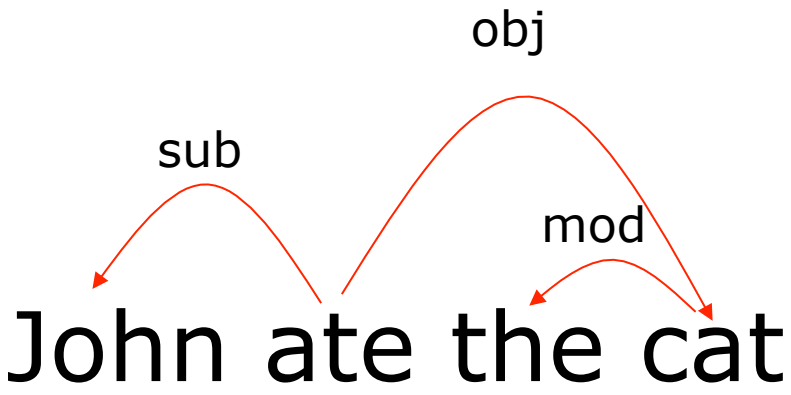
句法分析

- 组块分析（浅层句法分析、部分句法分析）
  - 基本短语（非递归的核心成分）识别
- 组成分分析（结构分析，完全句法分析）
  - 短语如何构成句子
- 依存分析
  - 词之间的依赖关系

"John ate the cat"的组成分分析



"John ate the cat"的依存分析



句法分析（Parsing）

- 句法分析的目的
  - 判断句子的合法性（句子识别）
  - 确定句子的结构（句子中单词相互关联的方式）
- 基于上下文无关语法（CFG）的表示
  - CFG能描述大部分的自然语言结构
  - 可以构造高效的基于CFG的句法分析器
- 通常采用树形结构来表示句法分析的结果

一个简单的产生式语法（英语）

1. S -> NP VP
2. VP -> V NP
3. NP -> NAME
4. NP -> ART N
5. NAME -> John
6. V -> ate
7. ART -> the
8. N -> cat
9. ……

- 产生式5～9属于词法规则，一般由词典与词性标注算法来描述

优秀语法的特征

- 通用性
  - 能正确分析句子的范围
- 选择性
  - 能判断出错误句子的范围
- 可理解性
  - 自身的简易程度
- *鲁棒性
  - 对不合法句子的容忍度：He love her.
  - 通用性与选择性矛盾的处置，如：忽略主谓一致性检查将导致无法区分下面句子的不同含义（歧义）
    - Flying planes are dangerous.
    - Flying planes is dangerous.

基于产生式的CFG分析器

- 自顶向下
  - 利用产生式，从S开始，尝试将S改写/推导成与输入句子相匹配的终结符号序列。
- 自底向上
  - 利用产生式，尝试将输入句子规约到S。
- 回溯
  - 在改写或规约的某一步可能有多个选择。
  - 从一个错误的尝试（改写或规约）返回，进行下一个尝试。
- 保留改写或规约的历史
  - 回溯需要
  - 输出正确的分析结果也需要

一个简单的自顶向下句法分析算法

- 语法
  - 1. S -> NP VP　　2. NP -> ART N　　3. NP -> ART ADJ N
  - 4. VP -> V　　5. VP -> V NP
- 位置计数器
  - $_1$ The $_2$ dogs $_3$ cried $_4$
- 状态
  - 由符号表和当前位置构成，如：((NP　VP)　1)　表示从位置1开始寻找NP，且NP后面是VP。
  - 分为当前状态和后备状态。
- 状态转换
  - 当前状态符号表的第一个符号是词法符号（词性），并且句子中当前词属于该词法类，则删除符号表中第一个符号，并更新当前位置(加1)，得到新的当前状态。
  - 当前状态符号表的第一个符号是句法符号，则依据语法获得改写该符号的所有产生式，把它们的右部作为符号表与当前位置构成状态；选择其中一个作为新的当前状态，其它作为后备状态（在回溯时使用）。
- 回溯
  - 从后备状态中取一个作为当前状态，继续分析

- 算法
  1. 取 ((S) 1)作为当前状态（初始状态），后备状态为空。
  2. 若当前状态为空，则失败，算法结束，
  3. 否则，若当前状态符号表为空，
     (1)当前位置处于句子末尾，则成功，算法结束
     (2)当前位置处于句子中间，转5
  4. 否则，进行**状态转换**，若转换成功，则转2
  5. 否则，**回溯**，转2。

“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP

步骤	*当前状态*	*后备状态*	*备注*
1	((S) 1)		初始状态
2	((NP VP) 1)		规则1改写
3	((ART N VP) 1)	((ART ADJ N VP) 1)	规则2、3改写
4	((N VP) 2)	((ART ADJ N VP) 1)	ART匹配the
5	((VP) 3)	((ART ADJ N VP) 1)	N匹配cat
6	((V) 3)	((V NP) 3) ((ART ADJ N VP) 1)	规则4、5改写
7	(() 4)	((V NP) 3) ((ART ADJ N VP) 1)	V匹配caught

“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程（续）

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP

步骤	*当前状态*	*后备状态*	*备注*
8	((V NP) 3)	((ART ADJ N VP) 1)	回溯
9	((NP) 4)	((ART ADJ N VP) 1)	V匹配caught
10	((ART N) 4)	((ART ADJ N) 4) ((ART ADJ N VP) 1)	规则2、3改写
11	((N) 5)	((ART ADJ N) 4) ((ART ADJ N VP) 1)	ART匹配a
12	(() 6)	((ART ADJ N) 4) ((ART ADJ N VP) 1)	N匹配mouse
13			结束

搜索策略

- 深度优先
  - 后备状态采用“栈”
  - 后备状态少，存储效率高
  - 面临“左递归”问题
- 广度优先
  - 后备状态采用“队列”
  - 后备状态多，存储效率不高

基于图的自底向上句法分析(chart parsing)

- 简单的自底向上句法分析效率不高，常常会重复尝试相同的匹配操作（回溯之前已匹配过）。
- 一种基于图的句法分析，其中，已经匹配过的结果被保存起来，今后需要时可直接使用它们，不必重新匹配。（动态规划）

Chart Parsing句法分析

- 图中结点表示句子中词之间的位置数字
- chart（非活动边）
  - 记录分析中规约成功所得到的所有词法和句法符号
- activearcs（活动边集）
  - 未完全匹配的产生式，用加小圆圈标记（°）的产生式来表示，如：
    - NP -> ART °ADJ N
    - NP -> ART °N
- agenda（待处理表）
  - 记录等待加入chart的匹配成功的词法和句法符号
- 上面的活动边、非活动边以及词法和句法符号都带有“始/终结点号”

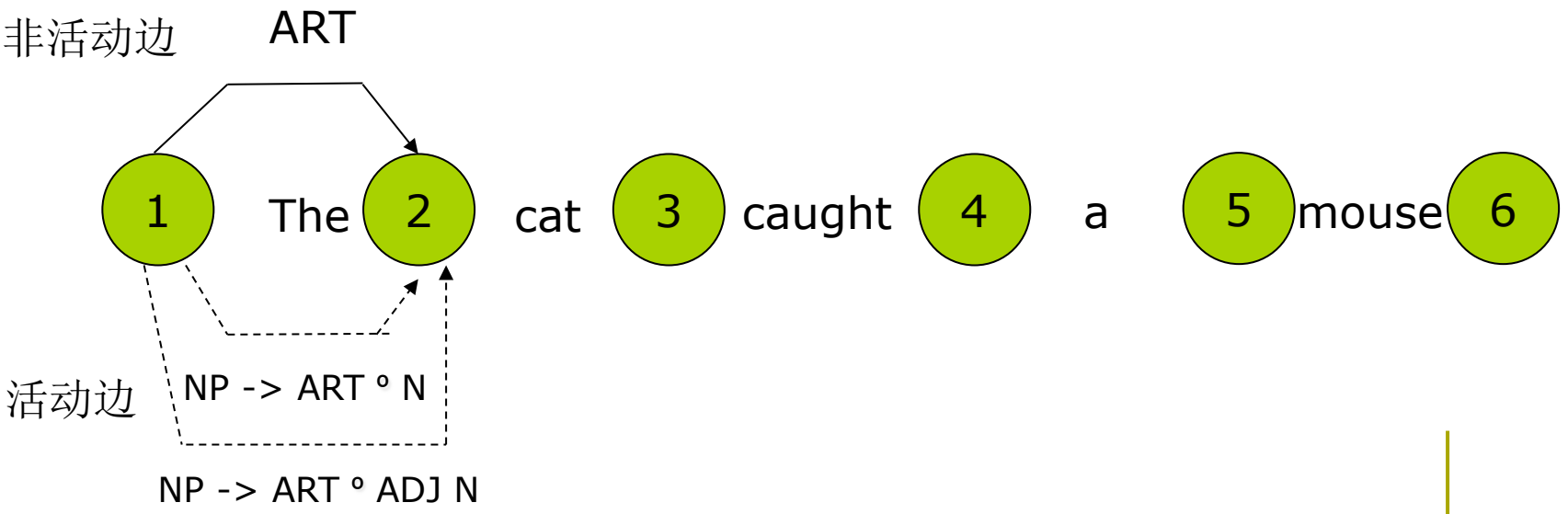
Chart Parsing句法分析算法（续）

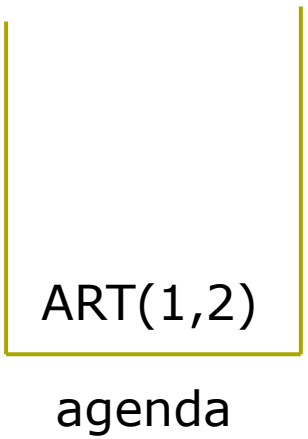
重复下面的操作直到agenda为空并且输入中没有下一个词

- 若agenda为空，则把句子中下一个词的各种词法符号（词性）加入进来，
- 从agenda中取一个元素（设为C，位置为：p1-p2)
- 对下面形式的每个规则：
  - X->CX_1...X_n，在activearcs中增加一条活动边：X->C ° X_1...X_n，位置为：p1-p2；
  - X->C，把X加入agenda，位置为：p1-p2
- 将C加入到chart的位置p1-p2
- 边扩展
  - 对每个形式为：X->X_1... ° C...X_n的活动边，若它在p0-p1之间，则在activearcs中增加一条活动边：X->X_1... C °...X_n，位置:p0-p2
  - 对每个形式为： X->X_1... X_n ° C的活动边，若它在p0-p1之间，则把X加入agenda ，位置为：p0-p2

1 The 2 cat 3 caught 4 a 5 mouse 6 的分析过程 (基法)

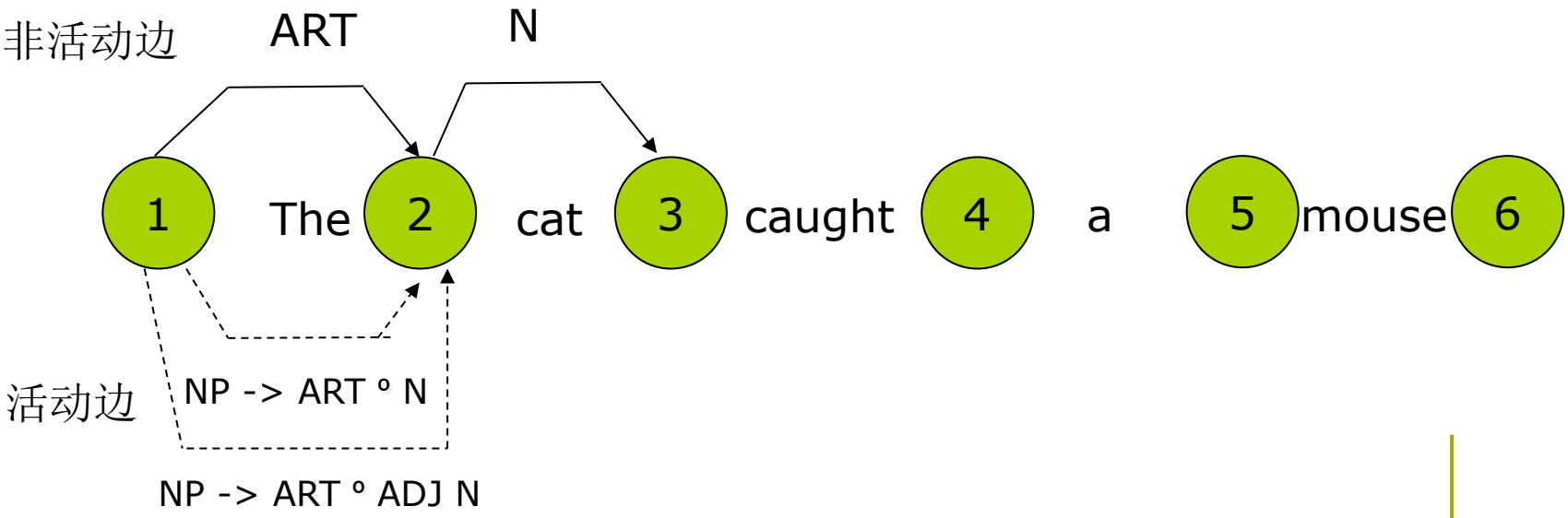
1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP





“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程 (算法)

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP

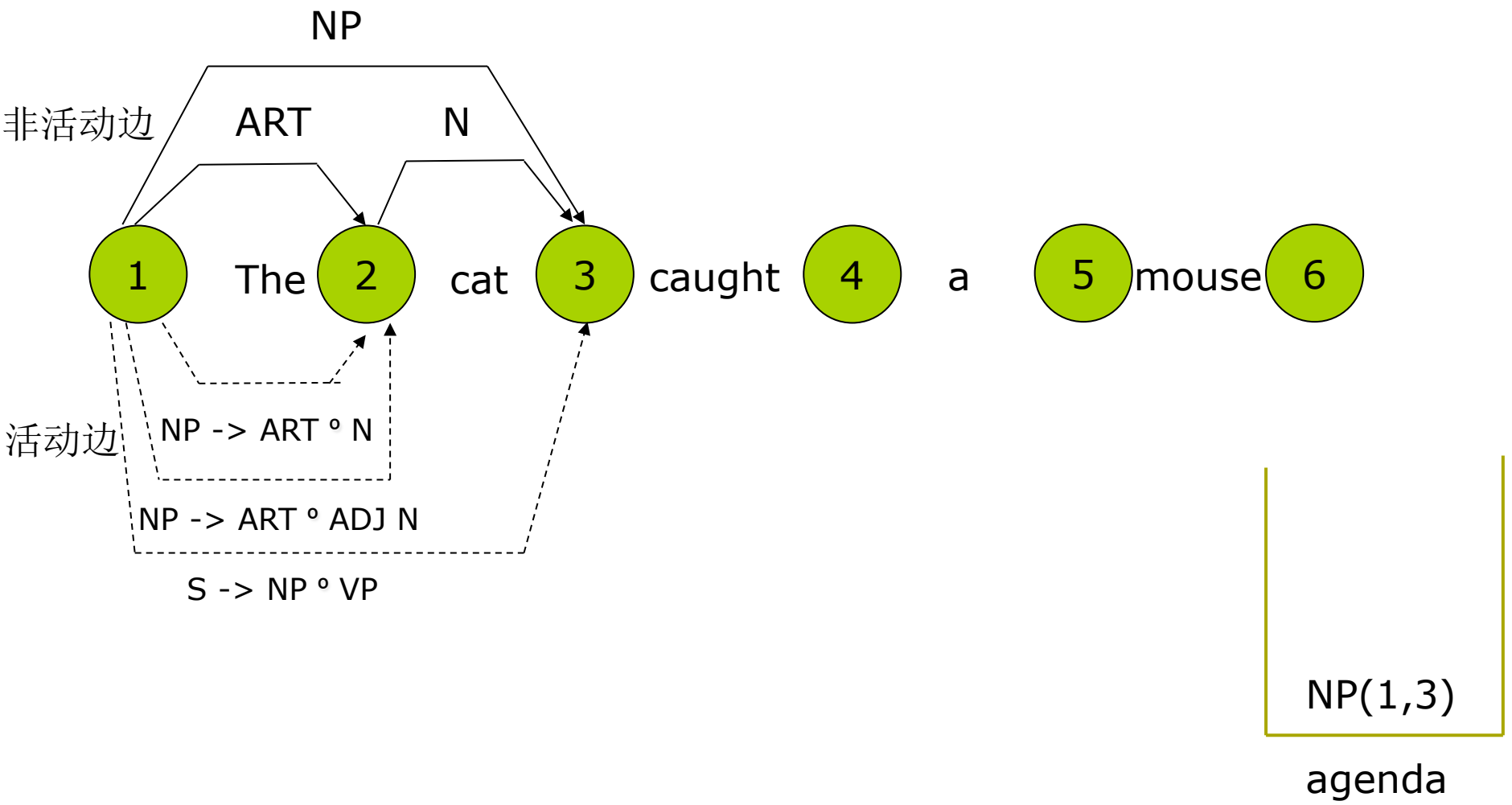


NP(1,3)

agenda

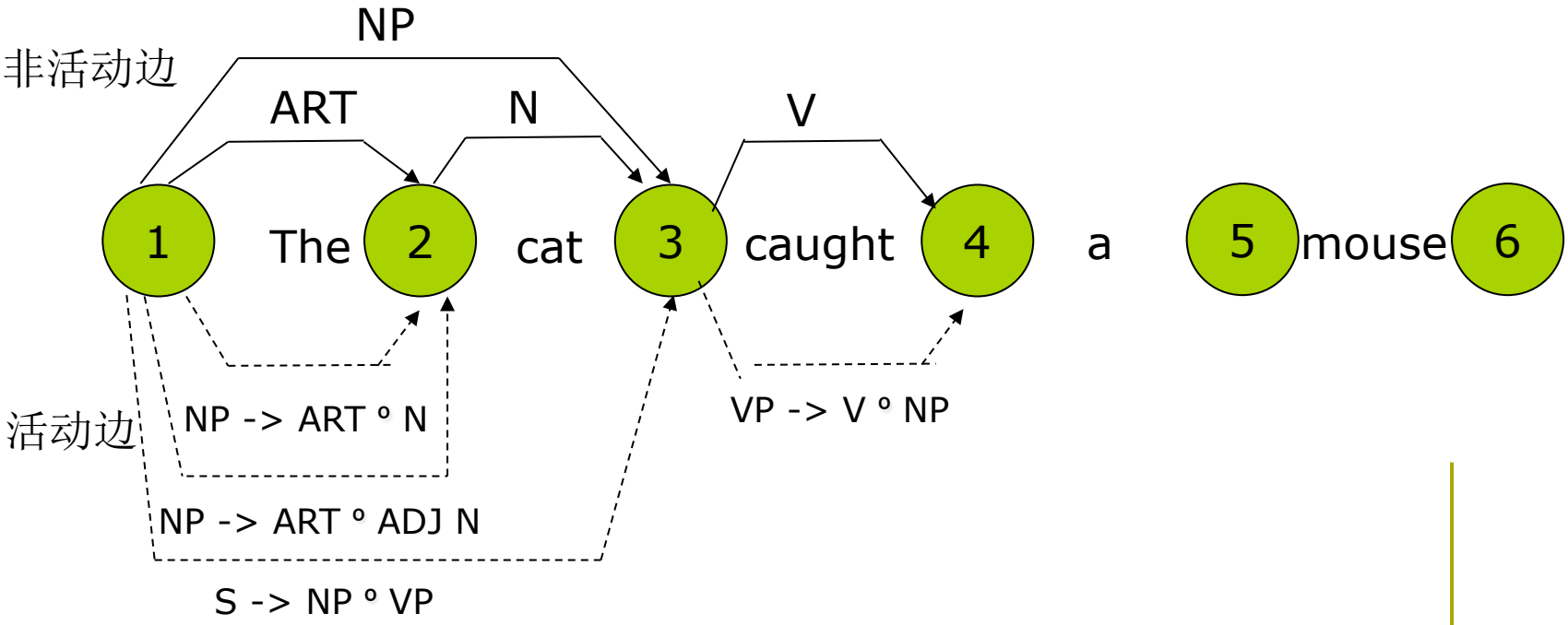
“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程 (算法)

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP



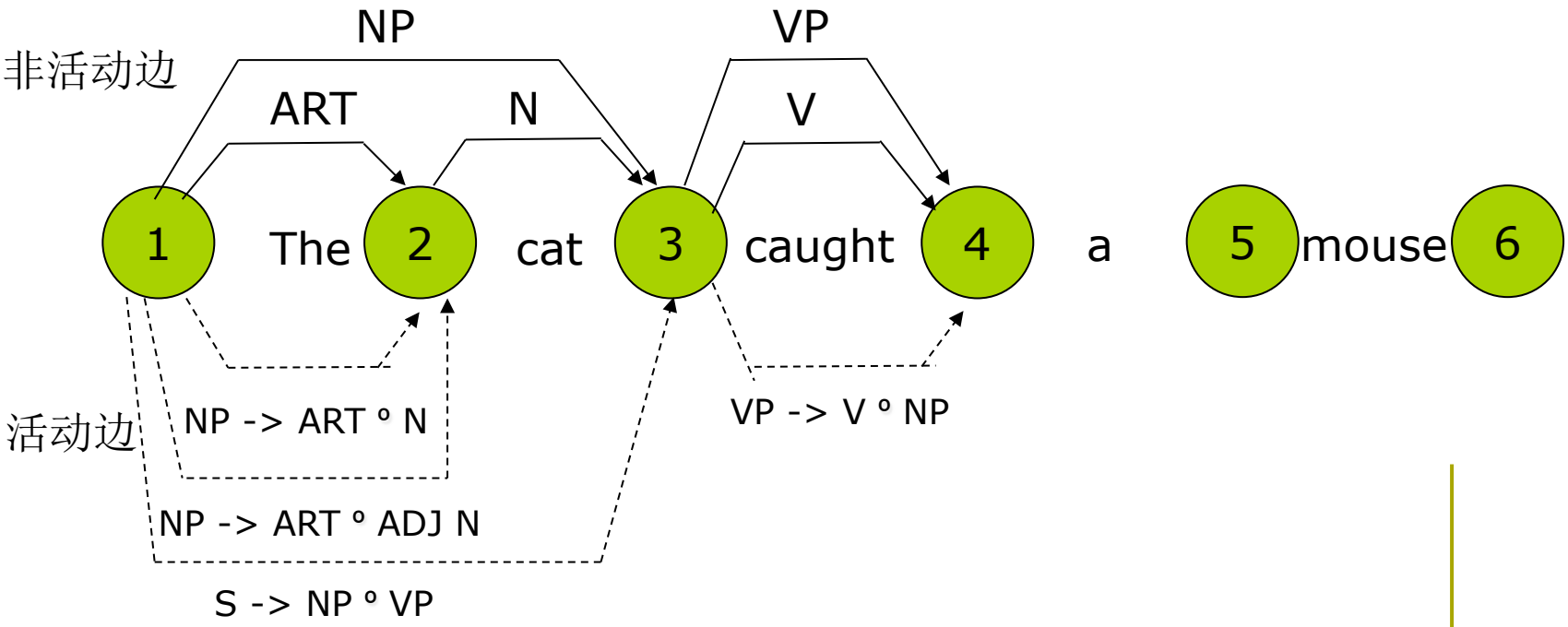
“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP



“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP

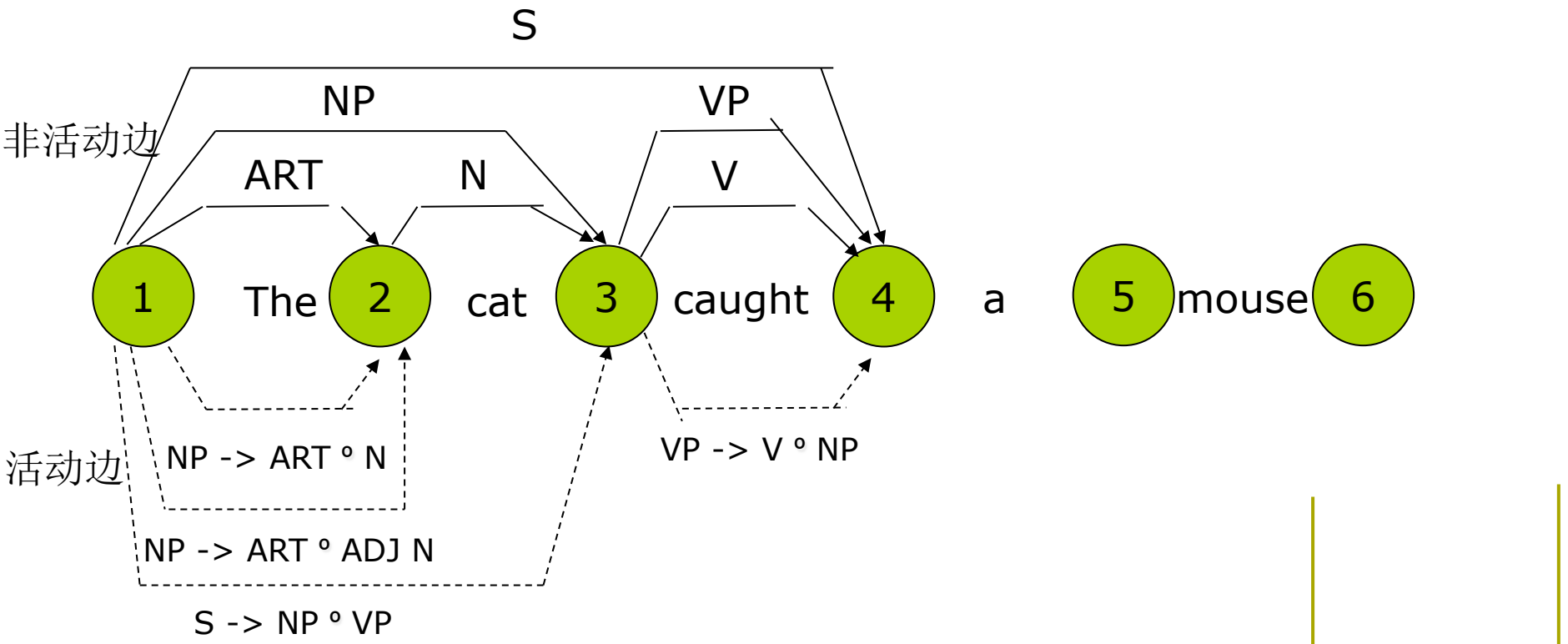


VP(3,4)

agenda

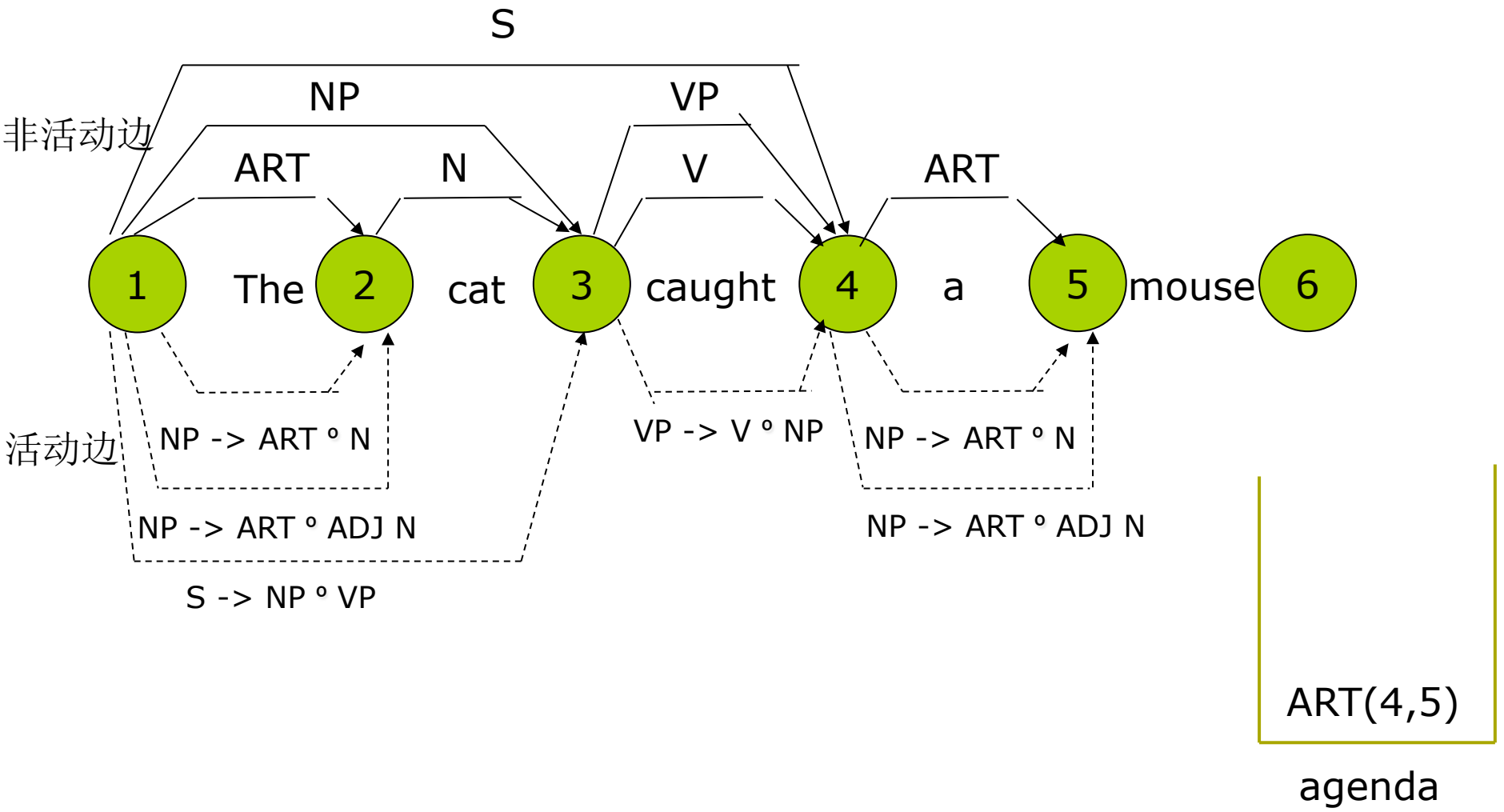
“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP



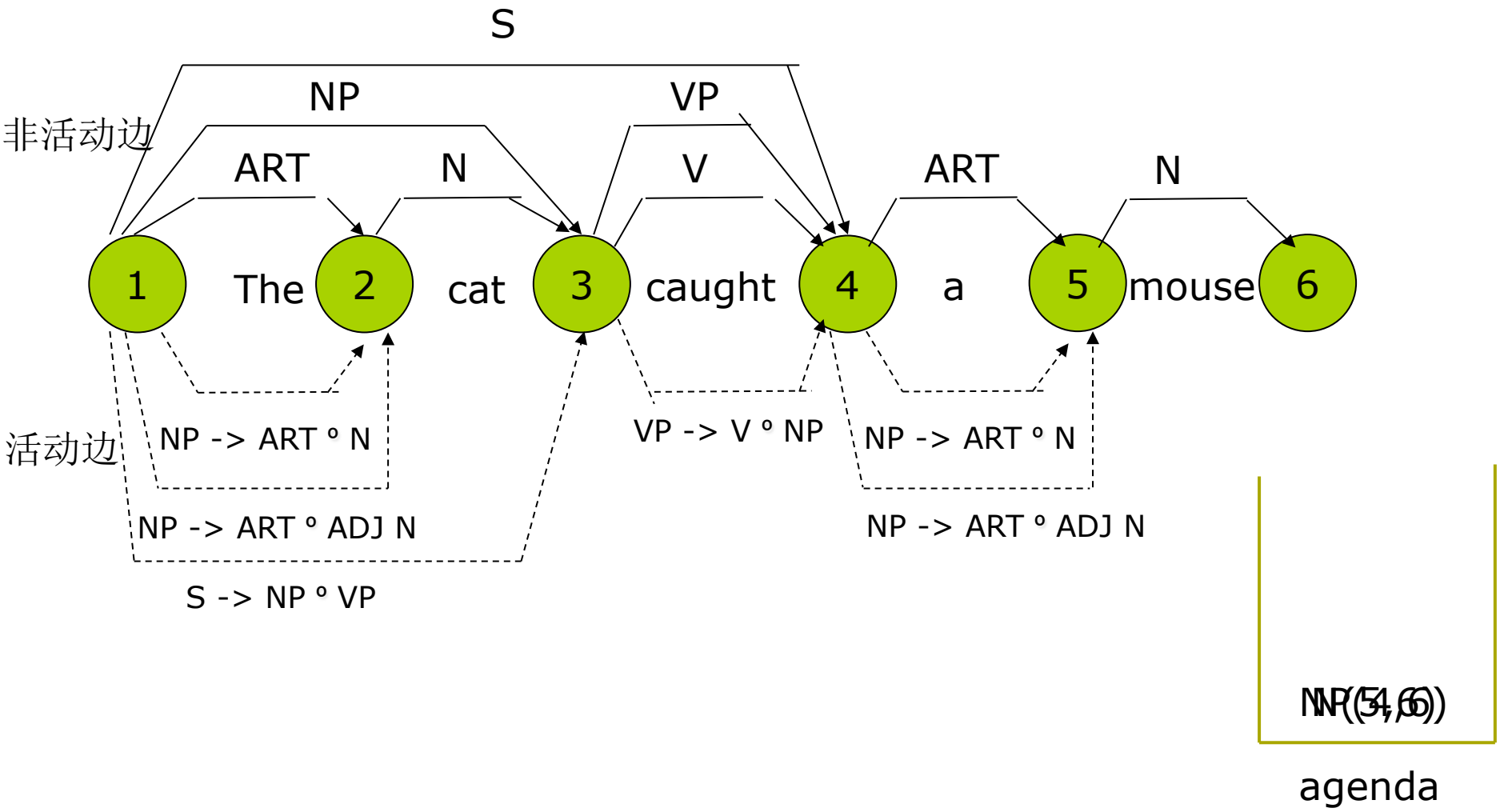
“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP



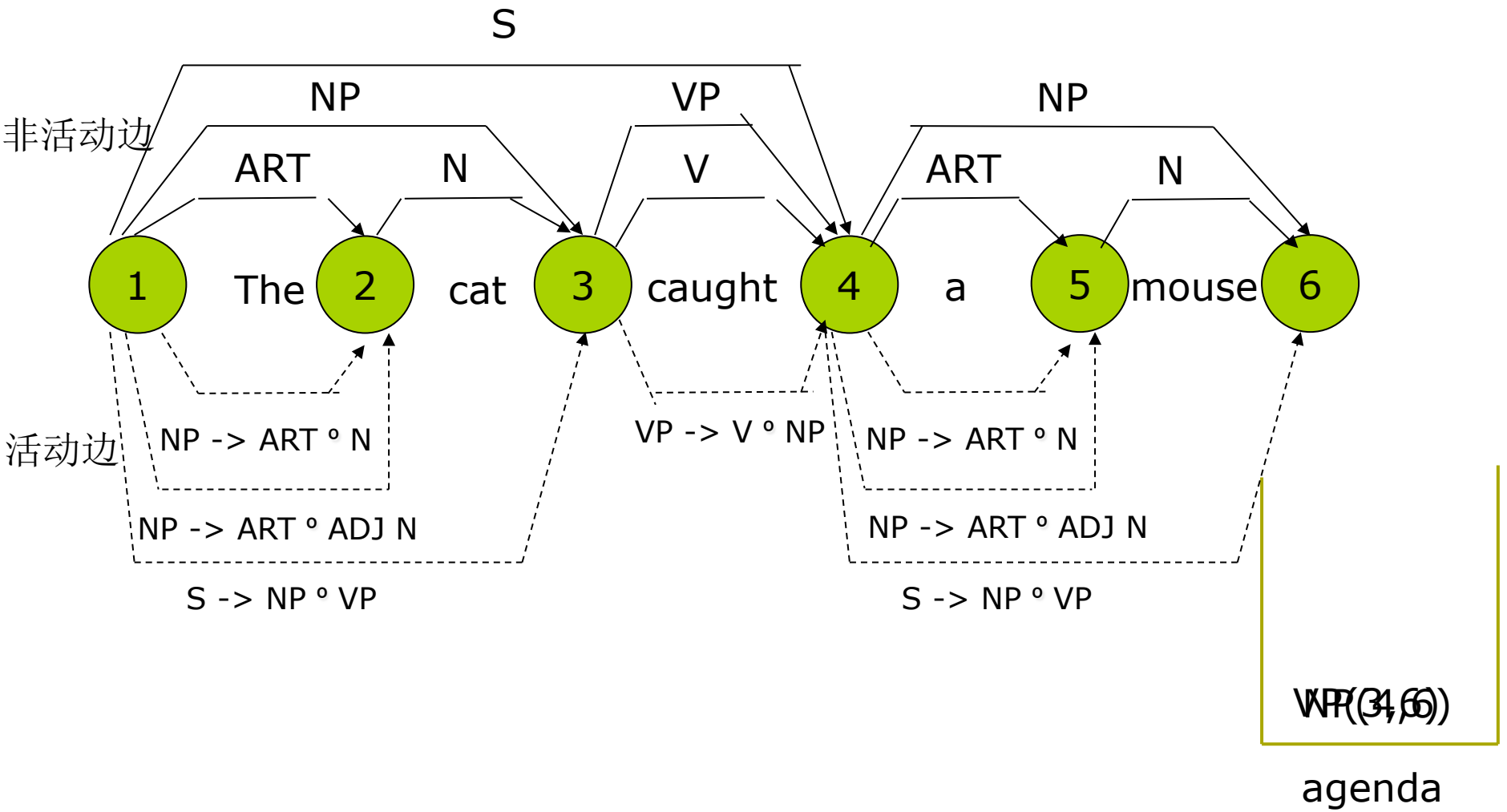
“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP



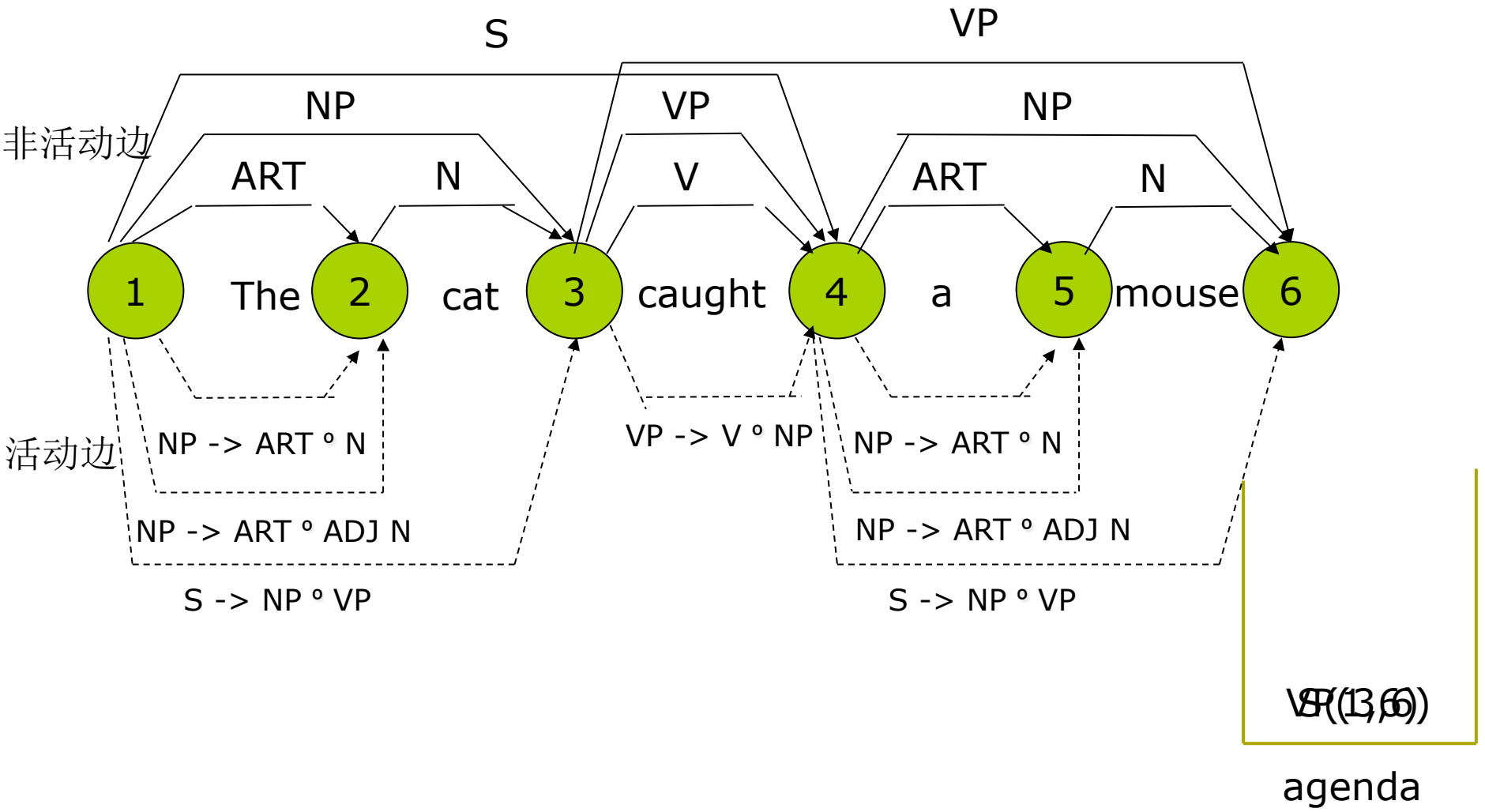
“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP



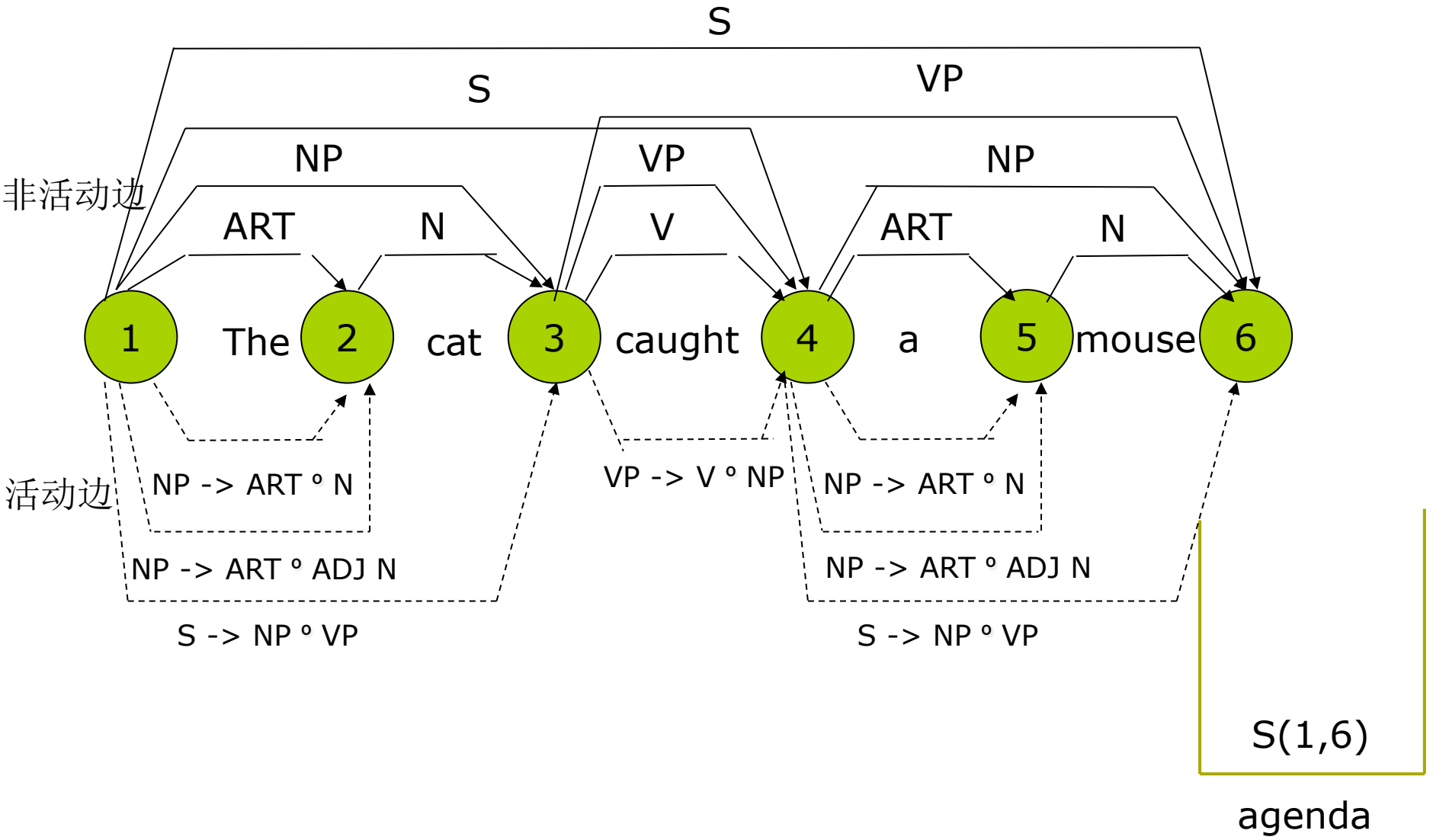
“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP



“1 The 2 cat 3 caught 4 a 5 mouse 6”的分析过程

1. S->NP VP 2. NP->ART N 3. NP->ART ADJ N 4. VP->V 5. VP->V NP



Proj. 3 实现一个基于简单英语语法的chart句法分析器。

句法分析与逻辑程序设计

- 逻辑程序设计是把程序组织成一组事实和一组推理规则，它基于谓词演算（Predicate Calculus）进行计算，计算过程由实现系统给出。
- 可以把语法写成PROLOG的子句（clause）：谓词（事实）和规则形式，推理过程由PROLOG的执行机制自动完成。

一个CFG语法的PROLOG表示

- 语法规则可表示成：
  - s(P1,P3):-np(P1,P2),vp(P2,P3)
  - np(P1,P3):-art(P1,P2),n(P2,P3)
  - np(P1,P3):-name(P1,P3)
  - pp(P1,P3):-p(P1,P2),np(P2,P3)
  - vp(P1,P2):-v(P1,P2)
  - vp(P1,P3):-v(P1,P2),np(P2,P3)
  - vp(P1,P3):-v(P1,P2),pp(P2,P3)
  - n(P1,P2):-word(W,P1,P2),isnoun(W)
  - art(P1,P2):-word(W,P1,P2),isart(W)
  - v(P1,P2):-word(W,P1,P2),isverb(W)
  - name(P1,P2):-word(W,P1,P2),isname(W)

- 词典可表示成：
  - isart(the)
  - isname(john)
  - isverb(ate)
  - isnoun(cat)
  - ……

- 输入句子 “John ate the cat”可表示成：
  - word(john,1,2)
  - word(ate,2,3)
  - word(the,3,4)
  - word(cat,4,5)
- 通过查询谓词s(1,5)的真假来识别句子 “John ate the cat”：
  - ?- s(1,5)
- 标准PROLOG的搜索策略与自顶向下的深度优先分析方法一致。

CFG在描述自然语言时存在的问题

1. S -> NP VP
2. NP -> ART N
3. NP -> ART ADJ N
4. VP -> V
5. VP -> V NP

- 上面的短语结构语法描述了英语的一个子集，同时，它又会生成一些不合法的英语句子，如：
  - The student solve the problem.（主谓不一致）
  - The teacher disappeared the problem.（不及物动词）

一种可能的解决方案——增加句法符号

- 把NP分为NP-S和NP-P；把VP分成VP-S和VP-P：
  - S->NP-S VP-S
  - S->NP-P VP-P
- 把N分成N-S和N-P：
  - NP-S->ART N-S
  - NP-S->ART ADJ N-S
  - NP-P->ART N-P
  - NP-P->ART ADJ N-P
- 把V分成V-S-I、V-S-T、V-P-I和V-P-T：
  - VP-S->V-S-I
  - VP-S->V-S-T NP-S
  - VP-S->V-S-T NP-P
  - VP-P->V-P-I
  - VP-P->V-P-T NP-S
  - VP-P->V-P-T NP-P

增加句法符号带来的问题

- 增加了规则的数量和潜在的冗余
- 类似的规则缺乏关联性
- 对语言结构描述缺乏深度（表层）

基于特征的扩展CFG

- 不增加原CFG中的句法符号
- 给每个句法符号增加特征（属性），例如：
  - NP(PER 3,NUM s)
  - VP(PER 3,NUM s,VAL itr)
- 特征由特征名和特征值构成。一系列特征构成了一个*特征结构*（*复杂特征集*）。
- 特征值可以是普通值（原子），也可以是另一个特征结构，例如：
  - NP(AGR （PER 3, NUM s))，可简写为：
  - NP(AGR 3s)
- 一个特征的特征值可以有多个，表示成：
  - N(ROOT fish, AGR {3s,3p})

- 特征值也可以是变量，例如：
  - NP(AGR ?a)
  - S->NP(AGR ?a) VP(AGR ?a) 表示NP与VP的AGR特征值一致（取同样的值）
- 一个规则如果包含特征值为变量的成分，则该规则代表了一组规则。（规则模板）
- 可以对变量形式的特征值限定范围（受限变量），例如：
  - NP(AGR ?a{3s,3p})

一个基于特征结构的CFG语法

- S->NP(AGR ?a) VP(AGR ?a)
- NP(AGR ?a) -> ART N(AGR ?a)
- NP(AGR ?a) -> ART ADJ N(AGR ?a)
- VP(AGR ?a) -> V(AGR ?a,VAL itr)
- VP(AGR ?a) -> V(AGR ?a,VAL tr) NP

基于合一的语法

- 一个文法可以表示成一系列特征结构间的约束关系所组成的集合。这样的文法称为合一文法（Unification Grammar）。例如：
  - 特征结构X0、X1和X2之间的约束关系：
    - X0->X1 X2 （CAT_0=S,CAT_1=NP,CAT_2=VP, AGR_0=AGR_1=AGR_2,$VFORM_0$=$VFORM_2$)
  - 它描述了基于特征的CFG中的一条规则：
    - S->NP(AGR ?a) VP(AGR ?a)
- 合一文法为其它的基于特征的文法提供了一个形式描述基础。
- 特征结构的合一运算构成了合一文法的基本操作，其作用有两个：
  - 创建新的特征结构（规约的结果）
  - 检查特征结构间的相容性以确定多个特征结构是否可以合并（规约）

合一运算

- 复杂特征集“相容”
  - α(f)表示复杂特征集α的特征f的值
  - 若α、β为复杂特征集，则α和β相容，当且仅当：
    - 若α(f)=a，β(f)=b，a、b都是原子，α和β是相容的当且仅当a=b
    - 若α(f)、β(f)均为复杂特征集，α和β是相容的当且仅当α(f)与β(f)相容
- 复杂特征集“合一运算”$\underline{\cup}$：
  - 如果a、b都是原子,若a=b，则a$\underline{\cup}$b=a，否则a$\underline{\cup}$b=Φ
  - 若α、β均为复杂特征集，则
    - 若α(f)=v，但β(f)未定义，则f=v属于α $\underline{\cup}$ β
    - 若β(f)=v，但α(f)未定义，则f=v属于α $\underline{\cup}$ β
    - 若α(f)=v_1，β(f)=v_2，且v_1与v_2相容，则f=($v_1\underline{\cup}v_2$)属于α $\underline{\cup}$ β，否则，α $\underline{\cup}$ β= Φ

合一运算举例

- (CAT V, ROOT cry)与(CAT V, VFORM pres)
  可以合一为：(CAT V, ROOT cry, VFORM pres)
- (CAT V, AGR 3s)与(CAT V, AGR 3p)不能合一
- (CAT N,ROOT fish, AGR {3s,3p})与(CAT N, AGR 3s)
  可以合一为： (CAT N,ROOT fish, AGR 3s)

基于特征CFG的chart parsing

- 句子语法成分与规则匹配时，要对各个特征进行匹配和泛化处理。
- 若规则包含特征值为变量的成分，匹配时需要实例化这个规则，例如：
  - 对于规则：
    - NP(AGR ?a)-> °ART(AGR ?a) N(AGR ?a)
  - 若有下面的语法成分需要匹配：
    - ART(ROOT a, AGR 3s)
  - 则需要实例化规则中的?a：
    - NP(AGR 3s)-> °ART(AGR 3s) N(AGR 3s)
  - 它与ART(ROOT a, AGR 3s)匹配后扩展为：
    - NP(AGR 3s)-> ART(AGR 3s) ° N(AGR 3s)
  - 若句子中还有N(ROOT dog, AGR 3s)需要匹配，则进一步扩展为：
    - NP(AGR 3s)-> ART(AGR 3s) N(AGR 3s) °

- 如果待匹配的语法成分的特征值中包含受限变量，则实例化后的规则中的取值范围为两者的交集，例如：
  - 实例化前的规则：
    - NP(AGR ?a)-> °ART(AGR ?a) N(AGR ?a)
  - 要匹配的语法成分：
    - ART(ROOT the, AGR ?a{3s,3p})
  - 实例化后的规则为：
    - NP(AGR ?a{3s,3p})-> °ART(AGR ?a{3s,3p}) N(AGR ?a{3s,3p})
  - 匹配扩展后为：
    - NP(AGR ?a{3s,3p})-> ART(AGR ?a{3s,3p}) °N(AGR ?a{3s,3p})
  - 再与N(ROOT dog, AGR 3s)匹配后扩展为：
    - NP(AGR 3s)-> ART(AGR 3s) N(AGR 3s) °

句义分析

- 句义分析的目的是给出句子的含义或意义(meaning)。句子的意义分为：
  - 上下文无关意义
  - 上下文有关意义
    - “Do you know what gate you are going to?”的意义是什么？
- 句义分析的作用：
  - 更好地进行翻译：Tom ran the machine.
  - 句法结构消歧：I saw a boy with a telescope.
- 句义分析的方式
  - 先句法后语义
  - 句法语义一体化
  - 完全语义分析（无句法分析）

词汇语义

- 句子的意义由句子中词汇的语义组合而成。句义分析首先需要解决词汇的语义表示和分析。
- 词汇的语义表示：
  - 义项（义位）
  - 语义类
  - 义素组合

义项（义位）

- 一个词往往有几个意义，每一个意义就是一个义项。例如：“明白”在《现代汉语词典》中的义项：
  - 内容、意义等使人容易了解；清楚；明确
  - 公开的、不含糊的
  - 聪明；懂道理
  - 知道；了解
- 义项之间的关系
  - 上下位关系：“动物”与 “狮子”
  - 整体-部分关系：“身体”与“上肢”
  - 同义关系：“美丽”与“漂亮”
  - 反义关系：“高”与“矮”
  - 包含关系：“兄弟”与“哥哥”和“弟弟”

- 表示义项之间关系的另一种方式是语义场——由几个相互关联的义项构成的语义系统。例如：
  - “师傅、徒弟”构成一个语义场
  - “上、下、左、右”也构成一个语义场
- 语义场的确定与本体论（Ontology）有关。贾彦德《汉语语义学》语义场的分类：
  - 分类义场：“中医、西医”、“城市、乡村”
  - 部分义场：“头、颈、躯干、四肢”、
  - 顺序义场：“优、良、及格、不及格”
  - 关系义场：“教师、学生”
  - 反义义场：“男人、女人”
  - 两级义场：“穷、富”、“大、小”
  - 部分否定：“必然、可能”
  - 同义义场：“警告、正告”、“掩饰、粉饰”
  - 枝干义场：“大、拍、捶”
  - 描绘义场：“白茫茫、白皑皑、白花花、白晃晃、白蒙蒙”

- 义项之间的关系可以为义项之间的搭配提供依据，从而为词义消歧（义项选择）和句义分析提供帮助。

语义类

- 由于义项的数量巨大，研究它们以及它们之间的关系非常困难。
- 解决这个问题的一种办法是：对义项进行泛化（抽象、概括）从而形成一些语义类（类似于词法分类——词性的做法）。例如：
  - 把“走”、“跑”、“跳”、“爬”几个义项泛化为语义类：“移动”。
  - 现代汉语词林
- 泛化的问题：
  - 语义类过多会失去泛化的效果。
  - 语义类过少会丢失信息。

义素（语义特征）

- 解决义项数量巨大的另一种方法是采用“义素”（语义特征）表示，义素是比义项更基本的语义单位。
- 一个义项可以表示成义素的集合（类似于句法中的复杂特征集）。例如：
  - “哥哥”的义素包括：“人、亲属、同胞、年长、男性”
- 在《知网》（http://www.keenage.com）中用“义原”表示。
- 义素为词汇语义提供了更精确的描述。

句义表示与分析（1）
——逻辑形式与语义组合

- 逻辑形式（LF，Logical Form）用于表示上下文无关的句义。它是对一阶谓词演算（FOPC）的扩充，增加了一些操作和广义量词。例如：
  - (DOG1 FIDO1)描述了句子：Fido is a dog.
  - (LOVES1 SUE1 JACK1)描述了句子：Sue loves jack.
  - (NOT (LOVES1 SUE1 JACK1))描述了句子：Sue does not love jack.
  - (MOST1 d1:(DOG1 d1)(BARKS1 d1))描述了句子：Most dogs bark.
  - (PRES(SEES1 JOHN1 FIDO1))描述了John sees Fido.
  - (EVERY b1:(BOY1 b1)(A d1:(DOG1 d1)(LOVES b1 d1)))描述了句子：Every boy loves a dog.的一个意思
  - (A d1:(DOG1 d1)(EVERY b1:(BOY1 b1) (LOVES b1 d1)))描述了句子：Every boy loves a dog.的另一个意思
  - (LOVES1 <EVERY b1(BOY1 b1)> <A d1(DOG1 d1)>)描述了句子：Every boy loves a dog.的两个意思（歧义表示）

- 语义组合：句子的语义由其成分的语义组合而成。
- λ演算为语义组合提供了形式化的计算基础和表示：
  - ((λxP(x))a) = P{x/a}
- 组合理论用于语义组合的难题：
  - 句法结构与逻辑形式之间存在结构上的不一致
  - 对习惯用语的处理（句义不由成分语义组合）
- 带语义解释的语法（语法/语义一体化）
  - 句法规则中加入语义特征，例如：
    - S(...,SEM (?semvp,?semnp))-> NP(...,SEM ?semnp)VP(...,SEM ?semvp)
  - 伴随句法规则给出句法符号的语义描述和计算规则

句义表示与分析（2）
——论旨角色与格语法

- 论旨角色（thematic role）或格角色（case role）
  - 基于动词给出句子中其它成分与它的浅层语义关系，例如：
    - The boy opened the door with a key.
      - the boy: AGENT（施事格）
      - the door: OBJECT（客体格）
      - a key: INSTUMENT（工具格）

格语法

- 格语法由美国语言学家Charles J. Fillmore提出的用于对句子的语义进行描述。（“Towards a modern theory of case”、“The case for case”、“Some problems for case grammar”）
- 基本语义规则
  - S->M+P
    - 一个句子（S）由情态（M）和命题（P）构成。
    - 情态包括：时体态、语气以及否定等。
  - P->V+C_1+C_2+...+C_n
    - 命题由动词（V）及若干格（C_1~C_n）构成。
  - C_i->K_i+NP_i
    - 格短语由格标记（K）和名词短语（NP）组成。
    - 提供从表层格到深层格的转换规则

- 格的种类：
  - 施事格(Agentive)：He laughed.
  - 工具格(Instrumental)：He cut the rope with a knife.
  - 与格(Dative)：He gives me a ball.
  - 使成格(Factitive)：John dreamed a dream about Mary.
  - 方位格(Locative)：He is in the house.
  - 客体格(Objective)：He bought a book.
  - 受益格(Benefective)：He sang a song for Mary.
  - 源点格(Source)：I bought a book from Mary.
  - 终点格(Goal)：I sold a car to Mary.
  - 伴随格(Comitative)：He sang a song with Mary.
  - ...(?)

- 动词格框架
  - 词典中对每个动词需给出：
    - 它所允许的格包括它们的性质（必需、禁止、自由）
    - 这些格的特征（附属词、中心词语义信息等）

基于格语法的语义分析

- 基于的信息
  - 格标记体系
  - 动词格框架
  - 名词语义信息
- 分析过程
  - 格短语及主动词识别
  - 利用主动词格框架确定格短语的格。
- 分析结果：句子的格框架。

基于格语法的语义分析结果（例）

- In the room, he broke a window with a hammer.

```
[BREAK
  [case-frame
      agentive: HE
      objective: WINDOW
      instrumental: HAMMER
      locative: ROOM]
  [modals
      time: past
      voice: active]
]
```

机器翻译

机器翻译历史

"I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need do is strip off the code in order to retrieve the information contained in the text"

- 1947，Warren Weaver's memo
- 1954，第一个公开展示的俄英MT原型系统
- 1966，美国科学院的ALPAC报告宣告机器翻译走入低谷
- 1970s，Systran（1970），Meteo（1976），
- Early 1980s，复苏，Eurotra，Mu
- Late 1980s~early 1990s，商品化系统投入市场，语音翻译，统计机器翻译（SMT）出现
- Late 1990s，Internet，MAT，EBMT
- After 2000，SMT大行其道！

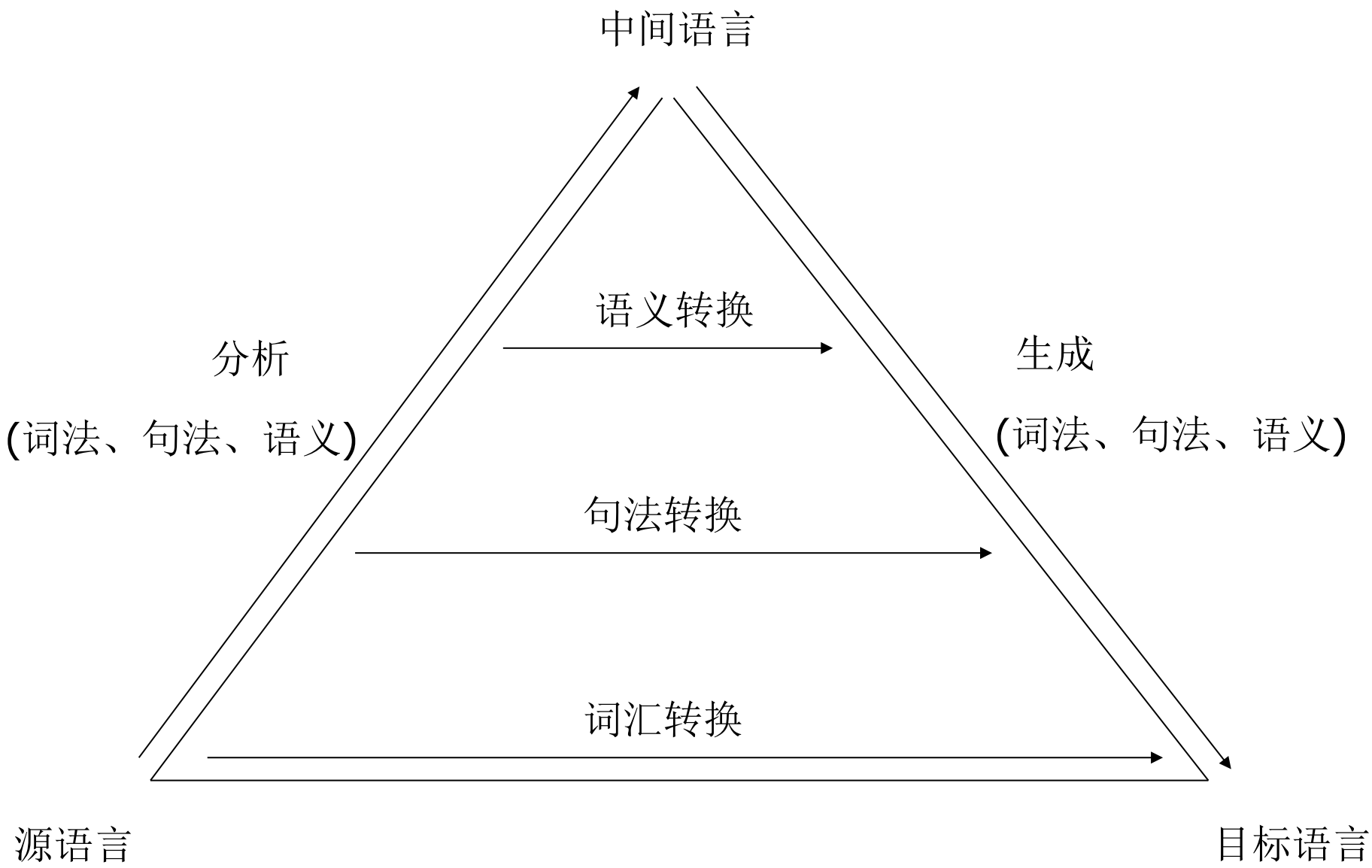
机器翻译的基本策略

- 直译（Direct）：从原文句子的表层（词、词组或短语）出发，直接转换成译文（必要的词序调整）。

- 转换（Transfer）：对源语言进行分析，得到一个基于源语言的中间表示；然后，把这个中间表示转换成基于目标语言的中间表示；从基于目标语言的中间表示生成目标语言。

- 中间语（Interlingua）：对源语言进行分析，得到一个独立于源语言和目标语言的、基于概念的中间表示；从这个中间表示生成目标语言。

机器翻译金字塔



机器翻译的实现方法

- 基于语言规则的理性方法（Rationalist approach）
  - 基于以规则形式表达的语言知识（词、句法、语义以及转换）进行推理。（Rule-based MT）
  - 又称传统的翻译方法，强调人对语言知识的理性整理。
  - Chomsky：先天语言能力，主宰1960－1985
- 基于语料库的经验方法（Empiricist approach）
  - 以大规模语料库（单语和双语）为语言知识基础。包括：
    - 基于统计的方法（SMT）
      - 利用统计学习方法自动获取和运用隐含在语料库中的知识
      - 翻译知识的获取在翻译之前完成，体现为一系列统计数据（参数）
    - 基于实例的方法（EBMT）
      - 基于类比原理，通过相似度计算，在语料库中找出最相似的句子
      - 翻译知识的获取在翻译之前没有全部完成,翻译过程中还需要语料库

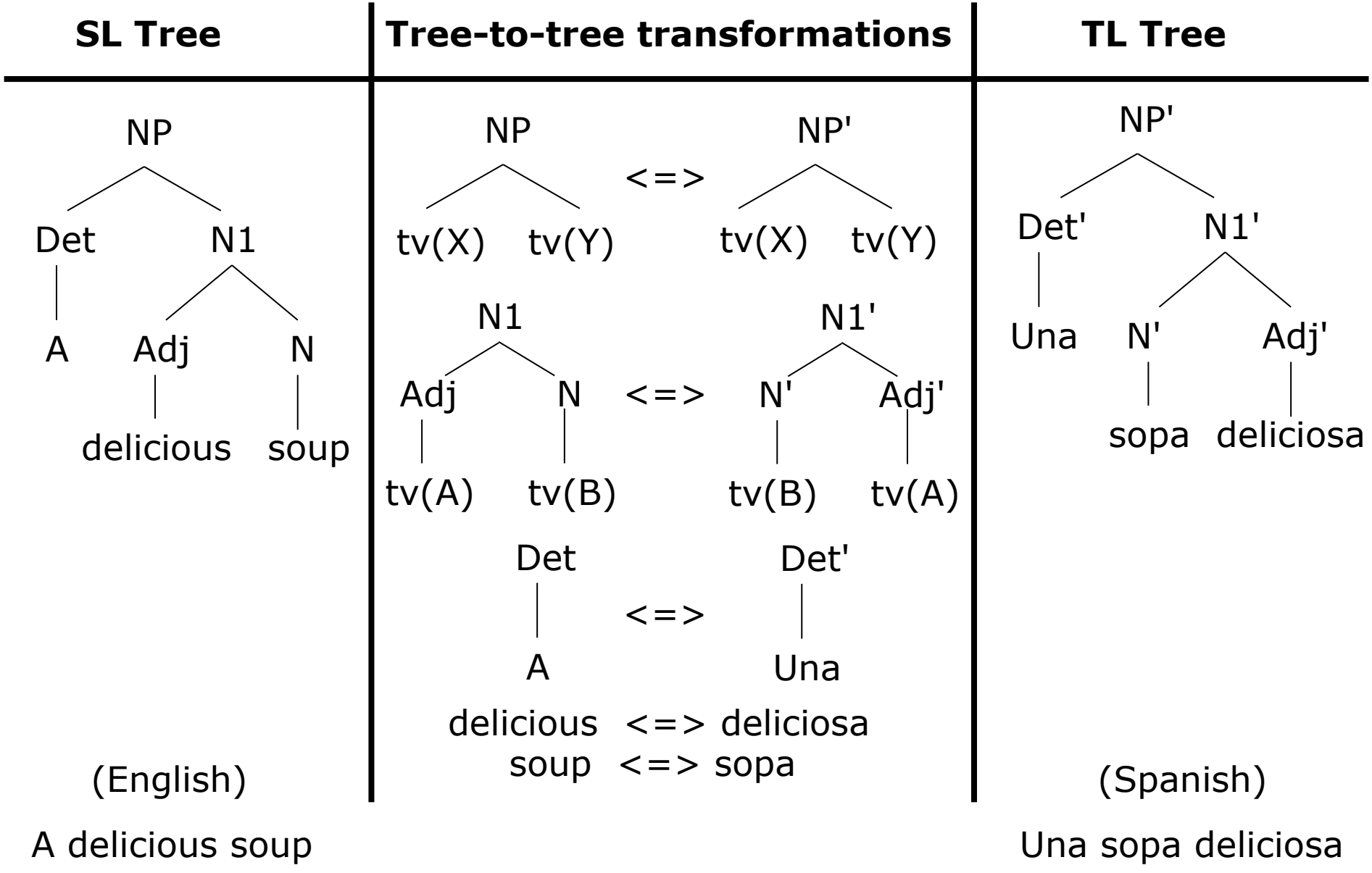
基于词的转换翻译

- 翻译过程
  - 译词选择
  - 词序调整
  - 形态（词形变化）生成
- 翻译所基于的知识
  - 对译（双语）词典及规则
  - 调序规则
  - 形态生成规则
- 问题
  - 没有句法结构和语义分析的指导，转换很难很好地进行，特别是对句法/语义结构相差很大的语言。
  - 译词选择和词序调整工作可用的信息太少（利用原句中的局部信息和已得到的译词信息） 。

基于句法结构转换的翻译

- 递归地利用一组“树-树”的转换规则，把源语言的分析树转换成目标语言分析树，然后生成目标语言句子。

句法树转换的例

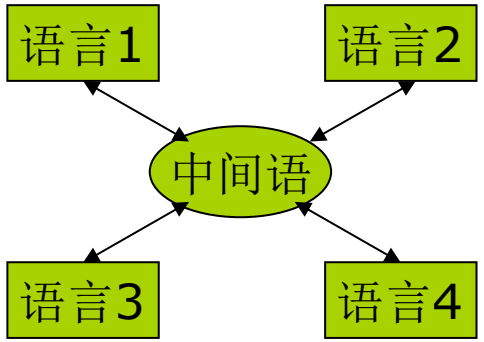

SL Tree
Tree-to-tree transformations
TL Tree
NP
Det
N1
A
Adj
N
delicious
soup
NP
tv(X) tv(Y)
<=>
NP'
tv(X) tv(Y)
N1
Adj
N
tv(A) tv(B)
<=>
N1'
N'
Adj'
tv(B) tv(A)
Det
A
<=>
Det'
Una
delicious <=> deliciosa
soup <=> sopa
NP'
Det'
N1'
Una
N'
Adj'
sopa
deliciosa
(English)
A delicious soup
(Spanish)
Una sopa deliciosa

基于语义转换的翻译

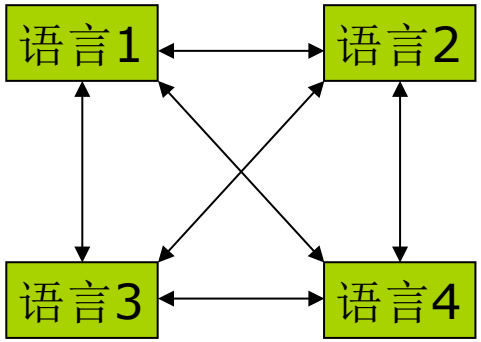
- 语义表示具有较高的语言独立性，在语义级转换避免了语言相关的句法结构转换。
- 转换规则需要解决不同语言之间的语义表示的对应问题：
  - 逻辑表示中的谓词转换
  - 论旨角色表示的格转换

基于中间语言(Interlingua)的翻译

- 基于中间语的翻译是指对源语言进行分析，得到一个独立于源语言和目标语言的、基于概念的中间语言表示，然后从这个中间语言表示生成目标语言。
- 对于n种语言之间的翻译（多语翻译）
  - 转换翻译需要n(n-1)个模块
  - 中间语言翻译需要2n个模块



中间语言翻译



转换翻译

- 中间语言翻译需解决的重要问题：一个统一的概念集及概念之间的关系集（本体论ontology所涉及的内容），使得它们对多种语言都适合。
- 中间语言翻译所需要的ontology是否存在？
- 中间语言翻译加大的语言分析的难度（大量的消歧）。（对机器翻译来说，这样的分析是否必要？）

机器翻译的现状

- 目前，机器翻译主要在一些简单的翻译任务中起到了一定的效果：
  - 对翻译质量要求不高的领域，如：网页浏览等
  - 子语言（领域受限）
  - 辅助翻译（后编辑）
- 任意文本的高质量的全自动翻译目前还很难实现。