

“自然语言处理导论”课程讲义

自然语言的序列标注问题 与解决方法(1)

孙栩

信息科学技术学院

xusun@pku.edu.cn

<http://klcl.pku.edu.cn/member/sunxu/index.htm>

- **链状结构即通常所说的“序列标注问题”**
- **自然语言处理的序列标注问题举例**
 - 词性标注
 - 中文切词
 - 短语识别（浅层句法分析）
 - 命名实体识别
 - ...
- **代表性的序列标注方法**
 - 关键问题是什么？
 - 隐马尔可夫模型 HMM
 - 结构化感知器 structured perceptron

□ 链状结构即通常所说的“序列标注问题”

□ 自然语言处理的序列标注问题举例

- 词性标注
- 中文切词
- 短语识别（浅层句法分析）
- 命名实体识别
- ...

□ 代表性的序列标注方法

- 关键问题是什么？
- 隐马尔可夫模型 HMM
- 结构化感知器 structured perceptron

例1：词性标注(Part-of-Speech Tagging)

- 给定一个句子(词序列)，对每个词标注出对应的词性类别
 - 即对每个词给出一个标签，即对每个词模式分类！在词性标注里，每个标签为一个词性(part-of-speech, POS)
 - 在句法分析、信息提取等任务上有重要作用

□ 英文词性标注举例:

定冠词 名词 动词过去式 介词

↓ ↙ ↘ ↙

DT NN VBD IN DT NN .

The cat sat on the mat .

词的分类依据

- 词类是一个语言学学术语，是一种语言中词的语法分类，是以语法特征（包括句法功能和形态变化）为主要依据、兼顾词汇意义对词进行划分的结果
- 词类划分具有层次性
 - 如汉语中，词可以分成实词和虚词，实词中又包括体词、谓词等，体词中又可以分出名词和代词等。

词的分类依据

□ 英语词类：分为10类

1. 介词 *preposition*
2. 定冠词 *determiner*
3. 代词 *pronoun*
4. 连词 *conjunction*
5. 名词 *nouns*
6. 动词 *verbs*
7. 形容词 *adjectives*
8. 副词 *adverbs*
9. 数词 *numeral*
10. 感叹词 *interjection*

词的分类依据

□ 词类的子类举例

□ 名词noun

1. 专属名词proper noun *eg. Beijing, IBM*
2. 通用名词common noun
 1. 可数名词countable noun *eg. book, table*
 2. 不可数名词mass noun *eg. communism, salt*

□ 副词adverb

1. 方向副词directional adverb *eg. downhill, home*
2. 程度副词degree adverb *eg. somewhat, extremely, very*
3. 方式副词manner adverb *eg. slowly, delicately*
4. 时间副词temporal adverb *eg. yesterday, tomorrow*

英语的总词类集合

- **宾州大学树库 *Penn treebank***
 - 45个词类
 - *Used for Penn treebank, WSJ Corpus*
- **布朗语料库 *Brown corpus***
 - 87个词类
 - *Used for Brown Corpus*

英语的总词类集合

宾州大学树库的部分词类举例

1.	CC	Coordinating conjunction	19.	PRP\$	Possessive pronoun
2.	CD	Cardinal number	20.	RB	Adverb
3.	DT	Determiner	21.	RBR	Adverb, comparative
4.	EX	Existential there	22.	RBS	Adverb, superlative
5.	FW	Foreign word	23.	RP	Particle
6.	IN	Preposition or subordinating conjunction	24.	SYM	Symbol
7.	JJ	Adjective	25.	TO	to
8.	JJR	Adjective, comparative	26.	UH	Interjection
9.	JJS	Adjective, superlative	27.	VB	Verb, base form
10.	LS	List item marker	28.	VBD	Verb, past tense
11.	MD	Modal	29.	VBG	Verb, gerund or present participle
12.	NN	Noun, singular or mass	30.	VBN	Verb, past participle
13.	NNS	Noun, plural	31.	VBP	Verb, non-3rd person singular present
14.	NNP	Proper noun, singular	32.	VBZ	Verb, 3rd person singular present
15.	NNPS	Proper noun, plural	33.	WDT	Wh-determiner
16.	PDT	Predeterminer	34.	WP	Wh-pronoun
17.	POS	Possessive ending	35.	WP\$	Possessive wh-pronoun
18.	PRP	Personal pronoun	36.	WRB	Wh-adverb

汉语中词的分类

□ 实词和虚词

- 从功能上看
 - 实词可以充当主语、谓语和宾语。 虚词则不可以
- 从意义上看
 - 实词有实在的意义，表示事物、动作、行为、变化、性质、状态、处所、时间等。 虚词基本只起语法作用，本身多无实在意义
- 从数量上看
 - 实词多为开放类。 虚词多为封闭类

□ 体词和谓词

- 实词通常可进一步分成体词和谓词
- 体词可以做主语和宾语。 谓词主要做谓语

汉语中词的分类

- 体词
 - 名词、处所词、方位词、时间词、区别词、数词、量词、代词
- 谓词
 - 动词、形容词
- 虚词
 - 副词、介词、连词、助词、语气词
- 拟声词、感叹词

兼类问题

- 如果同一个词具有不同词类的语法功能，则认为这个词兼属不同的词类，简称**兼类**
- 英文、中文都有兼类的问题，以中文为例
 - (1a) 共同完成任务 (1b) 共同愿望
 - (2a) 自动控制开关 (2b) 自动步枪
 - (3a) 定期检查机器 (3b) 定期存款
 - 在(a)组中，是副词
 - 在(b)组中是区别词

词类自动标注的任务

- 判定自然语言句子中的每个词的词类并给每个词赋以词类标记，例如：



难点

- 对于**兼类词**，词类标注程序应根据上下文确定兼类词在句子中最合适的词类标记

- **词类自动标注是深层语言分析的基础**

- 句法分析

- **词类标注程序判定依据**

- 要标注的词的不同词类的分布
- 上下文中其它词的词类信息
- 是个典型的序列标注问题

例2：中文切词(word segmentation)

- 通过计算机把组成汉语文本的字串自动转换为词串的过程被称为中文切词

- 即给定一个中文句子（字序列），尽可能将之切分成正确的词序列
- 是大部分中文信息处理任务的基础、第一步

- 例子

1: 跟前面的字切分

0: 跟前面的字不切分

1 0 1 1 0 1 0 1 0 1 1 0 0

企业要真正具有用工的自主权

结果：企业 / 要 / 真正 / 具有 / 用工 / 的 / 自主权

□ 自动的中文切词是许多应用的要求

□ 汉语切词是深层汉语分析的基础

- 句法分析
- 语义分析
- 信息检索

□ 语音处理

- 只有正确切词，才能知道正确的发音，如：
 - 的(de0) 目的(di4)

□ 问题

□ 切分歧义（消解）

- 一个字串有不只一种切分结果

□ 未登录词识别

- 专有名词
- 新词

切分歧义

□ 交集型歧义

- 从小/学/电脑
- 从/小学/毕业

□ 组合型歧义

- 美军/中将/竟公然说
- 新建地铁/中/将/禁止商业摊点

□ 混合型歧义

- 同时包含交集型歧义和组合型歧义的歧义字段

□ 未登录词识别

- 人名
- 地名
- 机构名
- 商标字号
- 专业术语
- 缩略语
- 新词语
 - 例如网络新词

未登录词识别困难

- 许多未登录词的构成单元本身都可以独立成词

切词的效果评价

□ 准确率(precision)

- 准确率 (P) = 切分结果中正确分词数/切分结果中所有分词数 *100%

□ 召回率(recall)

- 召回率 (R) = 切分结果中正确分词数/标准答案中所有分词数 *100%

□ F-指标 (F-score 综合准确率和召回率的评价指标)

- F-指标 = $2PR/(P+R)$

□ 基于规则的一些解决办法

□ 最大匹配法(MM)

- 1. 正向最大匹配法(MM)
- 2. 逆向最大匹配法(RMM)

□ 正向最大匹配法从左向右匹配词典

□ 逆向最大匹配法从右向左匹配词典

□ 例子

输入: 企业要真正具有用工的自主权

MM: 企业/要/真正/具有/用工/的/自主/权

RMM: 企业/要/真正/具有/用工/的/自/主权

以上规则方法的问题

- 切分算法需要有能力的检测到输入文本中何时出现了歧义切分现象
 - 以上的规则方法，包括MM和RMM法均没有检测歧义的能力
 - 只能给出一种切分结果
- 中文切词是个典型的序列标注问题

例3：短语切分 (phrase chunking)

- 给定一个自然语言的句子，对句子中的短语进行切分、并识别短语的种类
 - 又称为浅层句法分析(shallow parsing)
 - 对句法分析、机器翻译等任务有重要作用

□ 英文短语切分举例：

名词短语的开头

名词的继续

动词短语的开头

介词短语的开头

B-NP I-NP B-VP B-PP B-NP I-NP

The cat sat on the mat

例4：命名实体识别(named entity recognition)

- 给定一个句子或篇章，定位和识别相关的命名实体(named entity)
 - 命名实体包括：人名、地名、机构名
 - 或特定领域相关的命名实体，例如生物领域命名实体识别包括：蛋白质Protein、DNA、RNA等
 - 在信息提取、知识抽取等任务有重要作用

□ 举例

○ ○ ○ B-Protein I ○ B-DNA I I I

We showed that interleukin-1 IL-1 and IL-2 receptor alpha gene ...

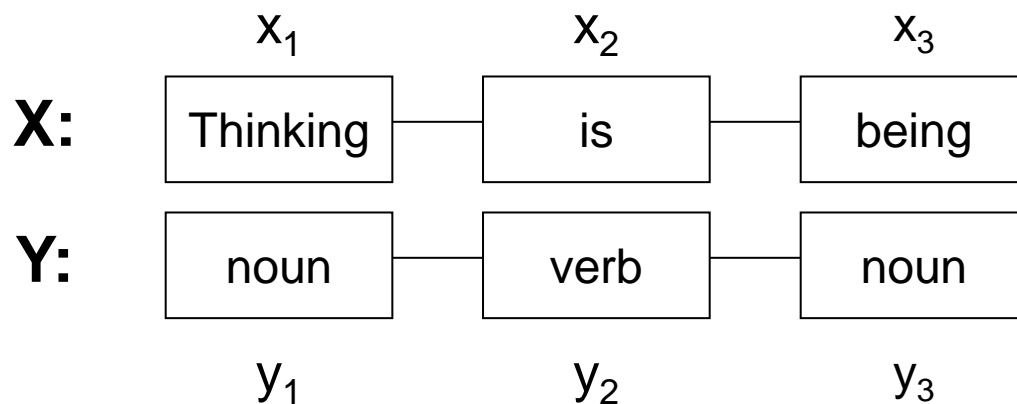
We showed that interleukin-1 IL-1 and IL-2 receptor alpha gene ...

Protein

DNA

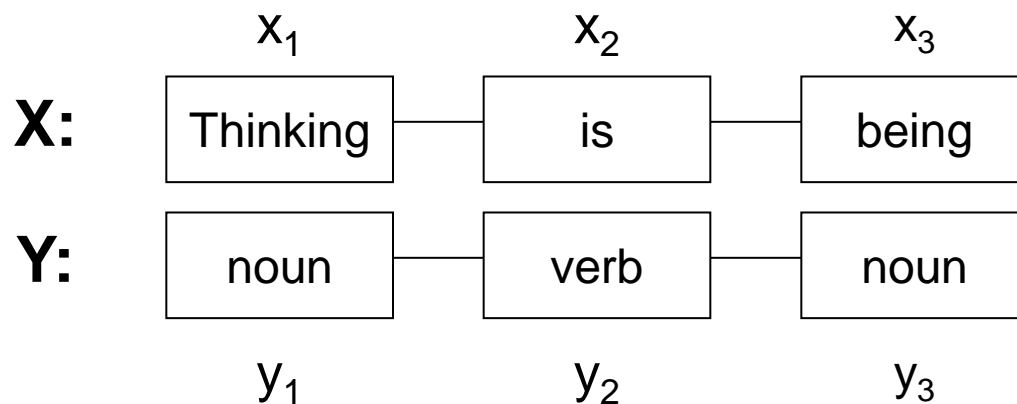
序列标注问题的总结

- 对一个序列中的每一个观测量(token)进行分类，给出一个对应的标签(label)
- 观测量对应的标签之间存在**结构依赖关系**，会互相影响，而不是独立分布的。这种结构依赖关系表现为较强的局部结构依赖，但是**局部结构依赖会传导为全局依赖**



序列标注问题的总结

- X 是一个观测向量，代表观测量序列
- Y 是一个标签向量，代表标签序列
- Y_i 是标签向量的第 i 个元素，值域是一个给定的有限集合：标签集合 A
- **序列标注问题：**
 - 给定一个有限标签集合 A ，学习怎么从观测向量 X 映射到标签向量 Y



□ 链状结构即通常所说的“序列标注问题”

□ 自然语言处理的序列标注问题举例

- 词性标注
- 中文切词
- 短语识别
- 命名实体识别

□ 代表性的序列标注方法

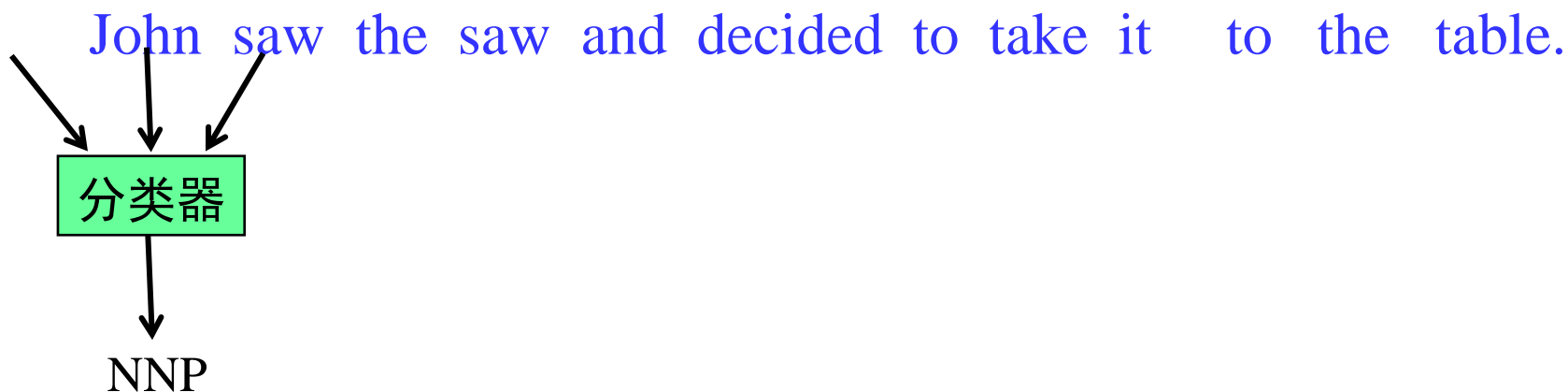
- 关键问题是什么？
- 隐马尔可夫模型 HMM
- 结构化感知器 structured perceptron

- **链状结构即通常所说的“序列标注问题”**
- **自然语言处理的序列标注问题举例**
 - 词性标注
 - 中文切词
 - 短语识别（浅层句法分析）
 - 命名实体识别
- **代表性的序列标注方法**
 - **关键问题是什么？**
 - 隐马尔可夫模型 HMM
 - 结构化感知器 structured perceptron

直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

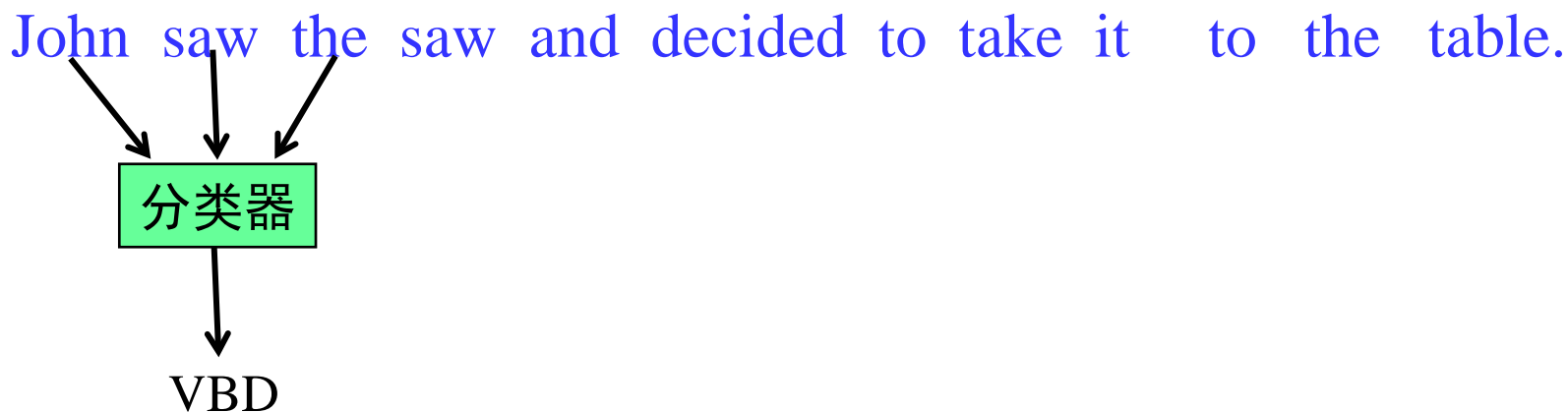
- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）



直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

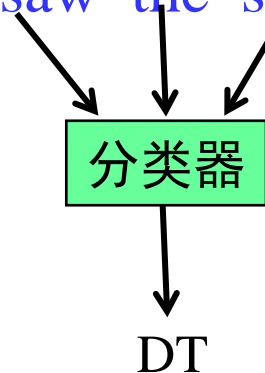


直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

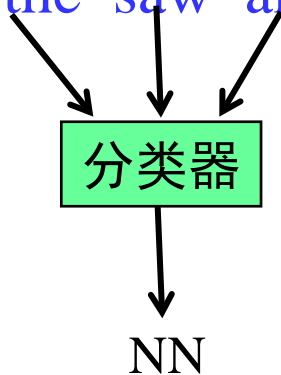


直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

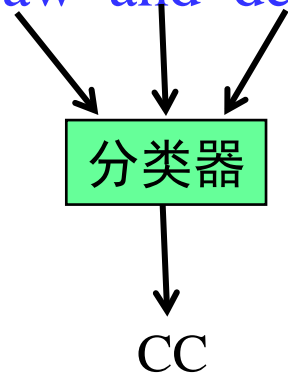


直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.



直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

分类器

The diagram illustrates a sliding window approach. Three arrows point from the words 'saw', 'and', and 'decided' in the sentence above to a green rectangular box labeled '分类器' (Classifier). This indicates that these three words are the input features for the classifier.

VBD

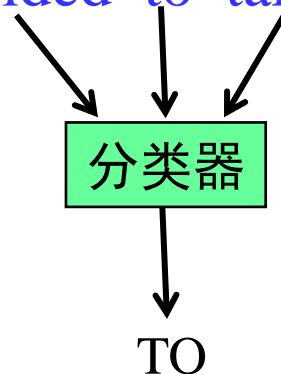
An arrow points from the '分类器' box down to the text 'VBD', representing the output of the classification process.

直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

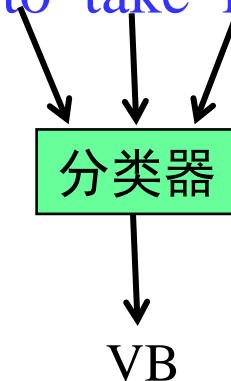


直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

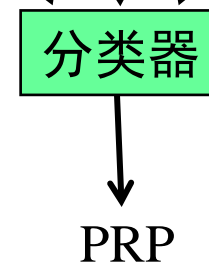


直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

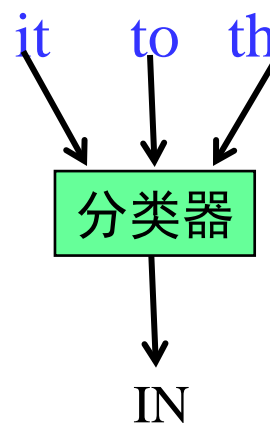


直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

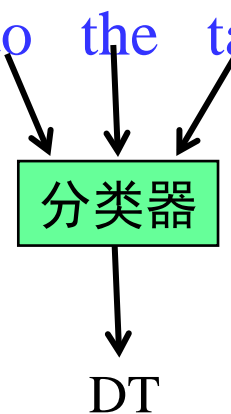


直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

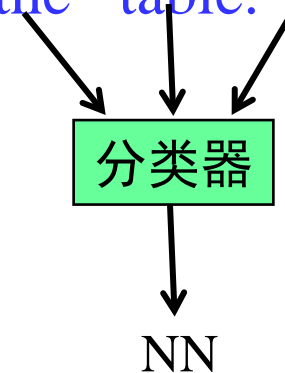


直接用简单分类方法会怎样？

□ 基于滑动窗口(sliding window)的简单分类方法

- 对每个观测量（词）进行独立的分类，使用周围的观测量（滑动窗口范围内的词）作为分类器的信息输入（提取的特征）

John saw the saw and decided to take it to the table.

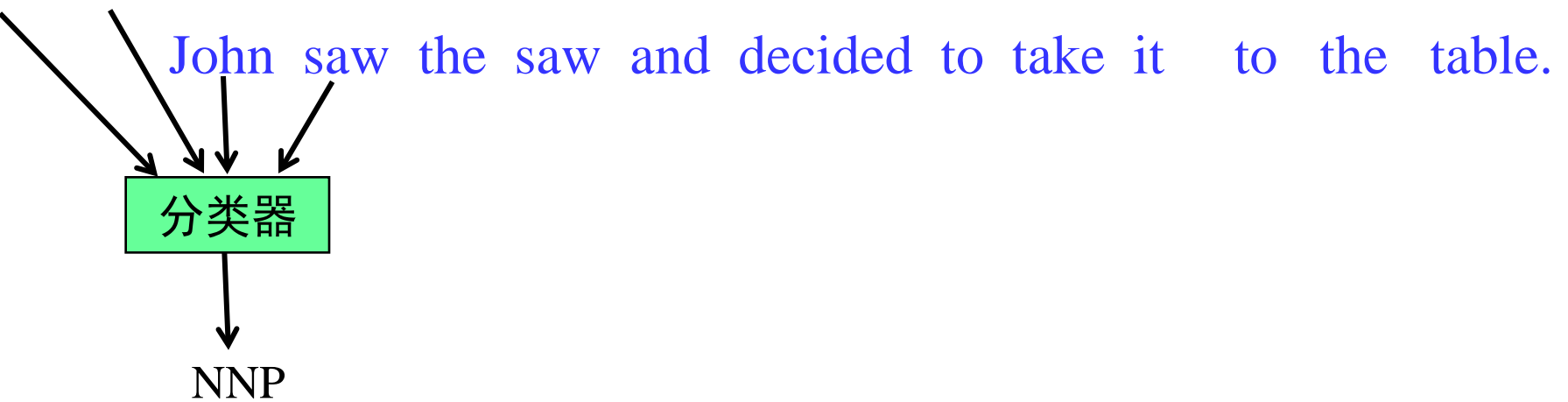


- 简单分类效果不好，因为无法考虑周围的标签信息（分类信息）

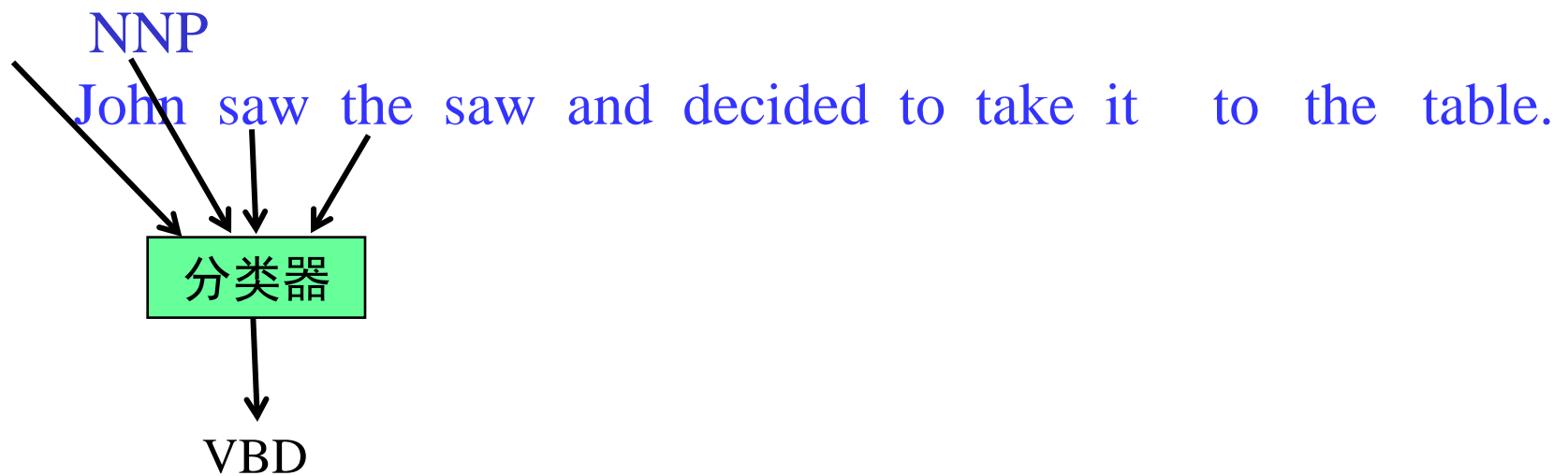
用改进的简单分类方法会怎样？

- 因为标签之间存在结构依赖关系，如果能够获得周围标签的信息，则能够对分类器形成更好的信息输入（即获得更好的特征）
- 问题是，周围的标签还不知道
- 一个解决办法是，可以采用前向或者后向的简单分类方法，获得周围的标签，从而改进原来的简单分类方法

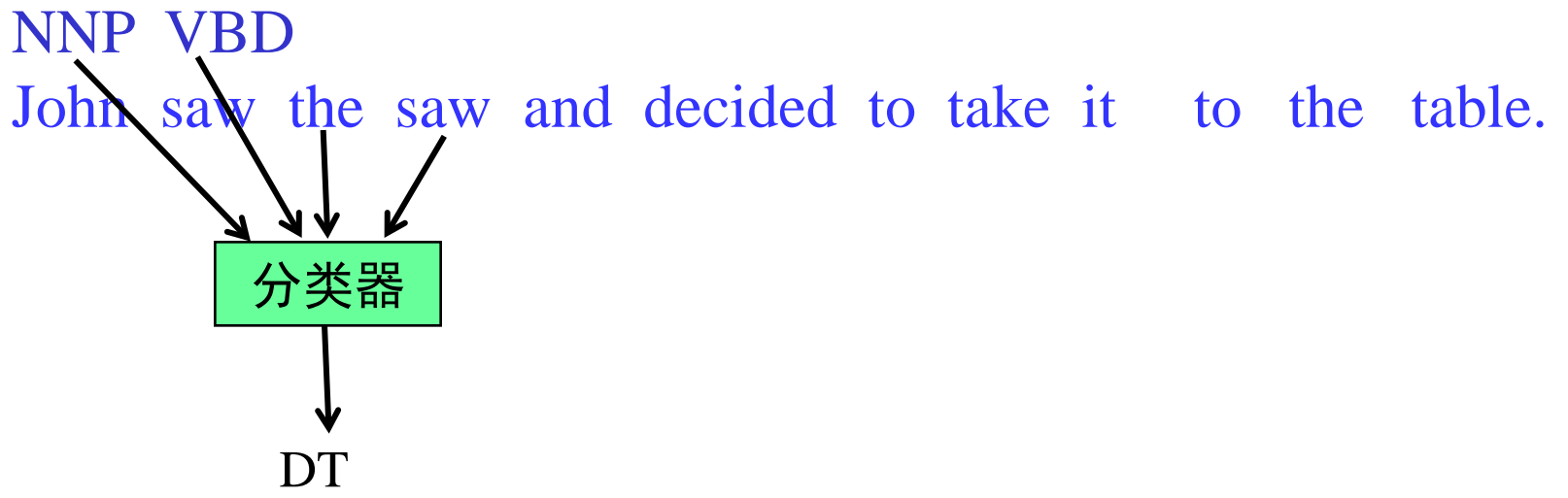
前向分类(Forward Classification)



前向分类(Forward Classification)



前向分类(Forward Classification)

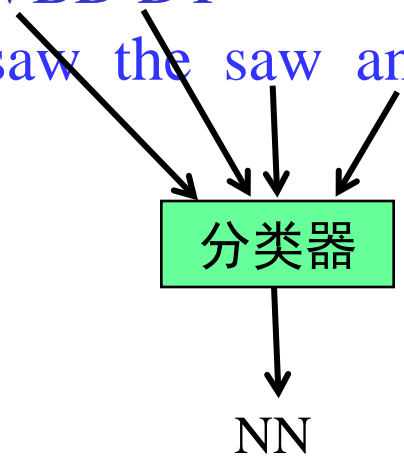


前向分类(Forward Classification)

NNP VBD DT

John saw the saw and decided to take it to the table.

分类器

A diagram illustrating the forward classification process. It shows a sequence of words: "John saw the saw and decided to take it to the table." Above the words "saw", "the", "saw", and "and" are the part-of-speech tags "NNP", "VBD", "DT", and "NNP" respectively. Arrows point from each of these four words to a central green box labeled "分类器" (Classifier). An arrow points from the classifier box down to the text "NN".

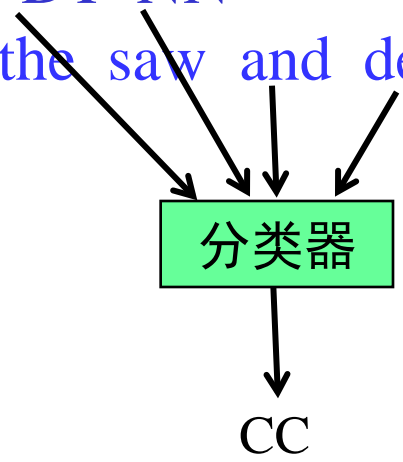
NN

前向分类(Forward Classification)

NNP VBD DT NN

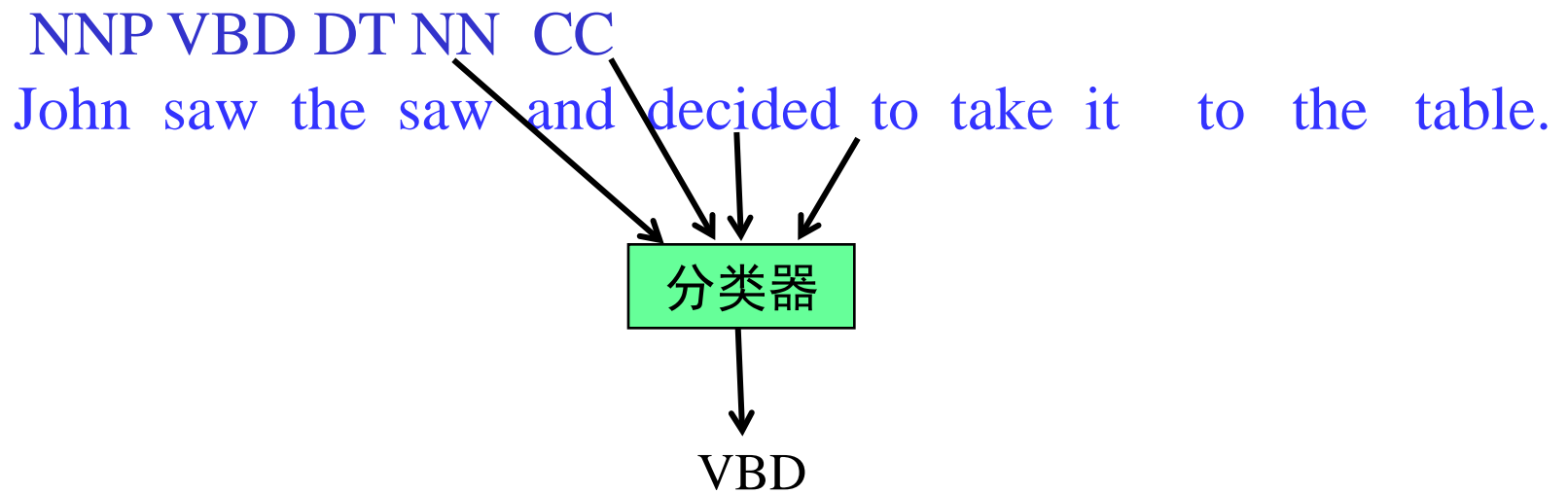
John saw the saw and decided to take it to the table.

分类器

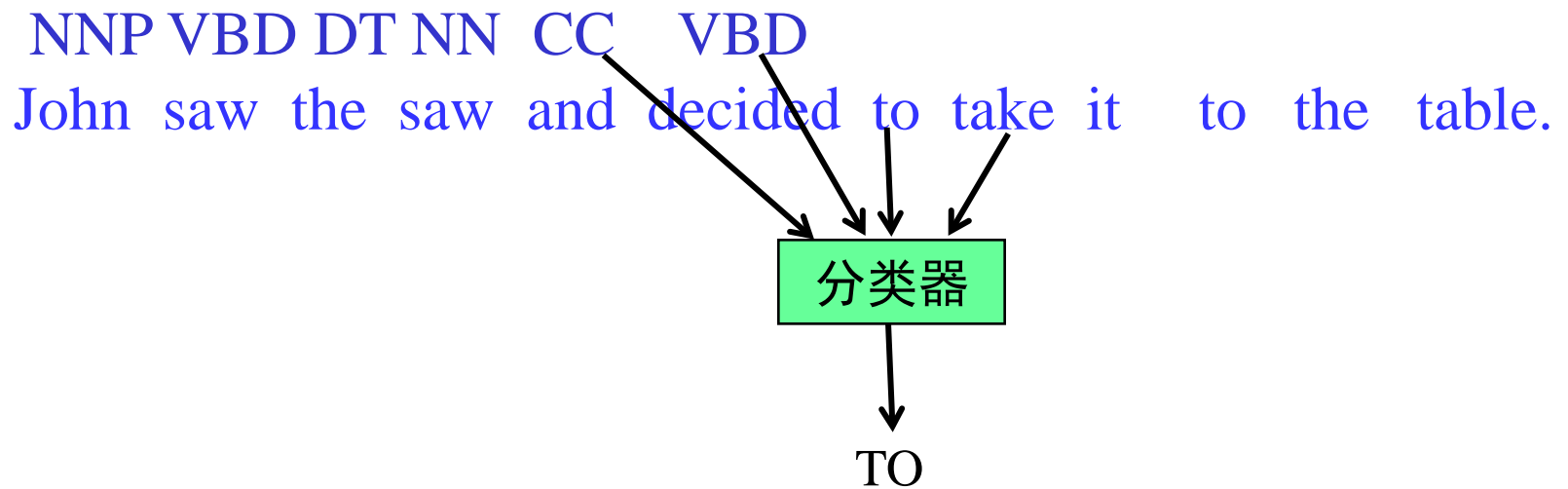
A diagram illustrating the forward classification process. It shows a sequence of words: "John saw the saw and decided to take it to the table." Above the words "saw", "saw", "and", and "decided" are the part-of-speech tags "VBD", "DT", "NN", and "VBD" respectively. A green box labeled "分类器" (Classifier) is positioned below the words "saw", "saw", "and", and "decided". Four arrows point from the words "saw", "saw", "and", and "decided" to the classifier box. An arrow points from the classifier box to the tag "CC" below it.

CC

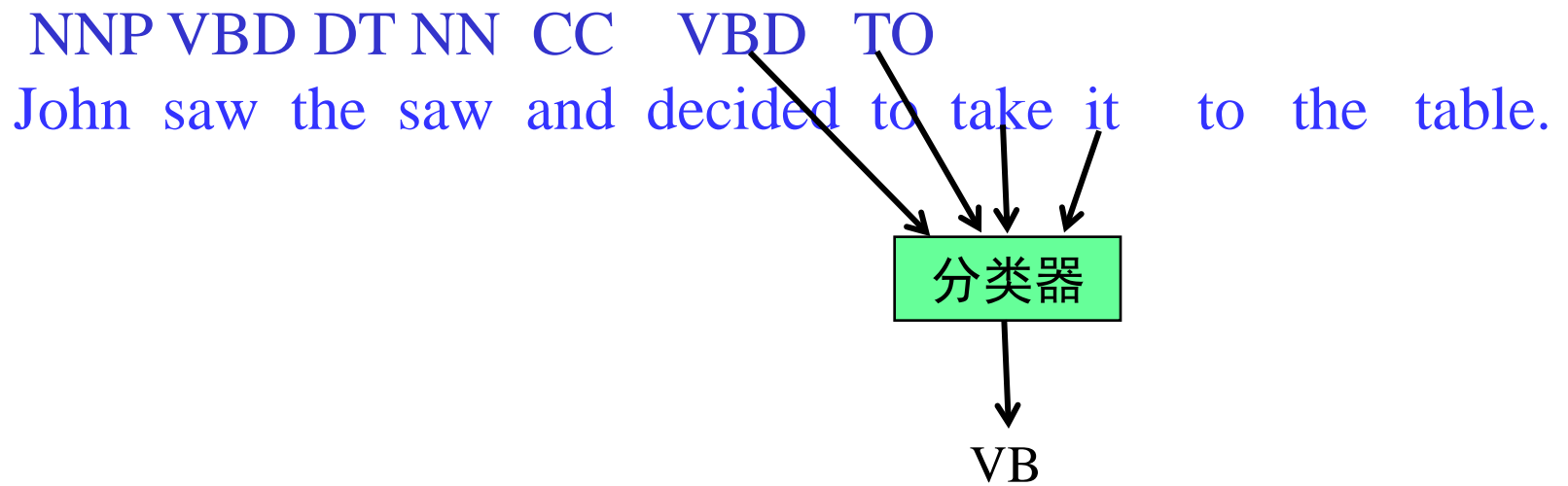
前向分类(Forward Classification)



前向分类(Forward Classification)



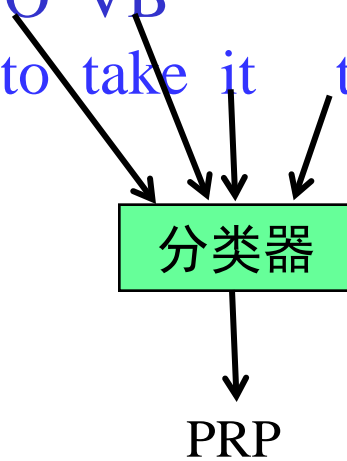
前向分类(Forward Classification)



前向分类(Forward Classification)

NNP VBD DT NN CC VBD TO VB
John saw the saw and decided to take it to the table.

分类器

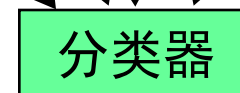
A diagram illustrating the forward classification process. It shows a sequence of words from a sentence: "John saw the saw and decided to take it to the table." Above the words, their corresponding Part-of-Speech (POS) tags are listed: "NNP VBD DT NN CC VBD TO VB". Arrows point from the words "to", "take", "it", and "to" to a green rectangular box labeled "分类器" (Classifier). An arrow then points from the classifier box down to the POS tag "PRP", which is the predicted class for the word "it".

PRP

前向分类(Forward Classification)

NNP VBD DT NN CC VBD TO VB PRP
John saw the saw and decided to take it to the table.

分类器

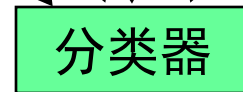
A green rectangular box with a black border containing the Chinese characters "分类器" (Classifier). Four black arrows point from the words "to", "take", "to", and "the" in the sentence above to the top edge of the box. A single black arrow points from the bottom center of the box to the word "IN" below it.

IN

前向分类(Forward Classification)

NNP VBD DT NN CC VBD TO VB PRP IN
John saw the saw and decided to take it to the table.

分类器

A green rectangular box with a black border containing the Chinese characters '分类器' (Classifier). Four black arrows point from the words 'to', 'the', 'table', and 'table.' in the sentence above to the top edge of the box.

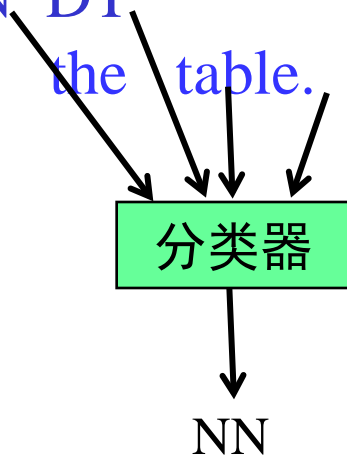
DT

A black arrow points from the bottom center of the '分类器' box to the text 'DT' below it.

前向分类(Forward Classification)

NNP VBD DT NN CC VBD TO VB PRP IN DT
John saw the saw and decided to take it to the table.

分类器



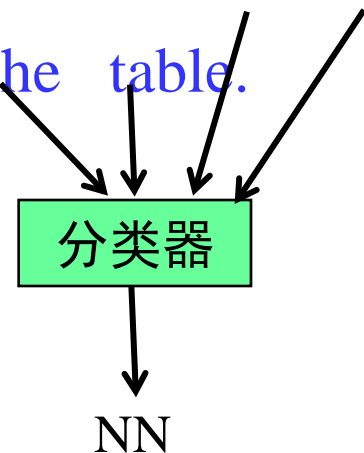
```
graph TD; A1[to] --> B[分类器]; A2[the] --> B; A3[table] --> B; B --> C[NN];
```

NN

后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势

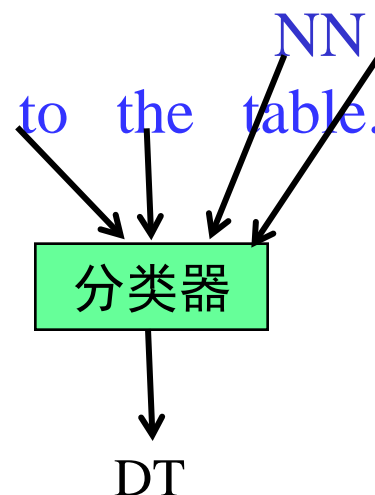
John saw the saw and decided to take it to the table.



后向分类(Backward Classification)

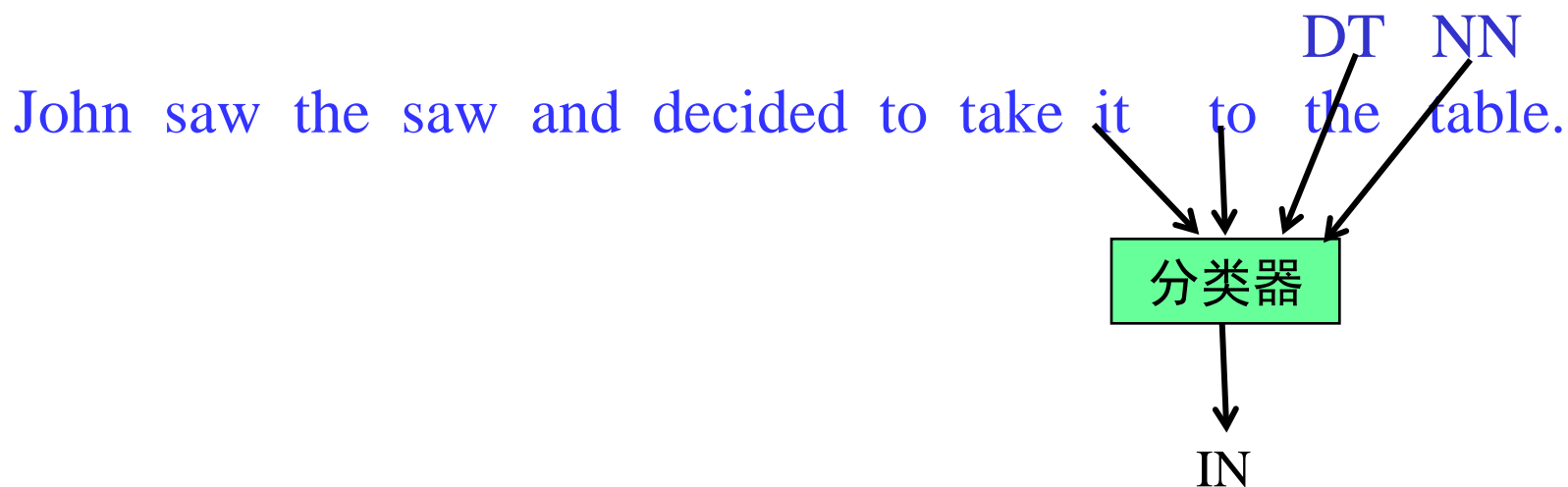
- 对 “to” 进行分类的时候，后向算法比前向算法有优势

John saw the saw and decided to take it to the table.



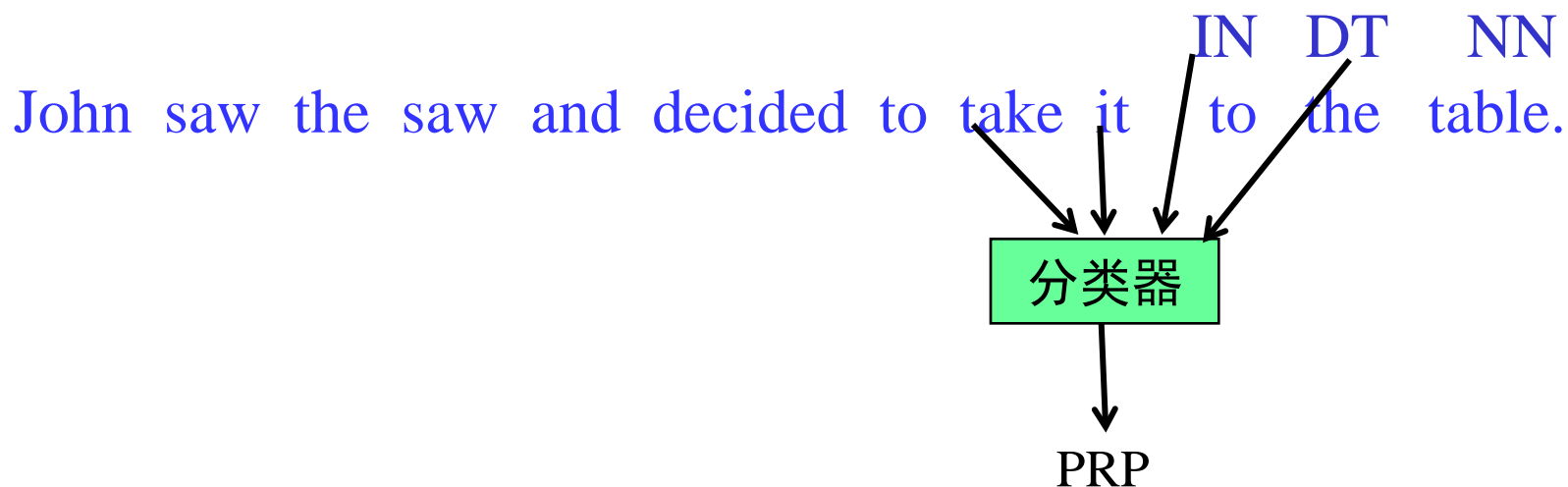
后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势



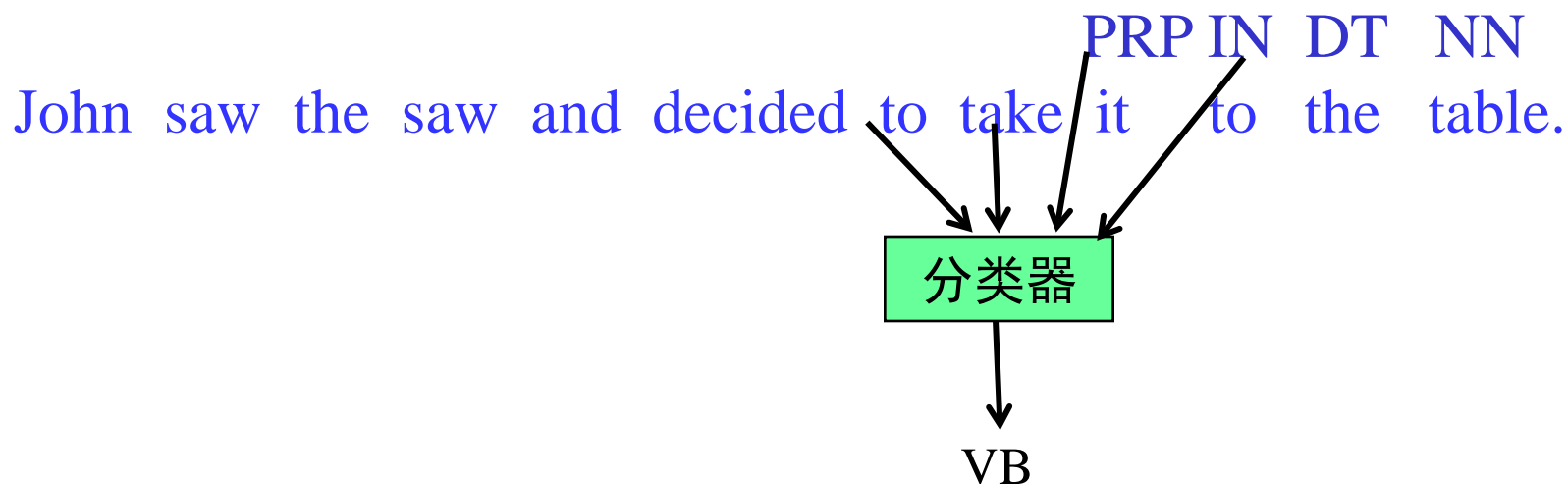
后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势



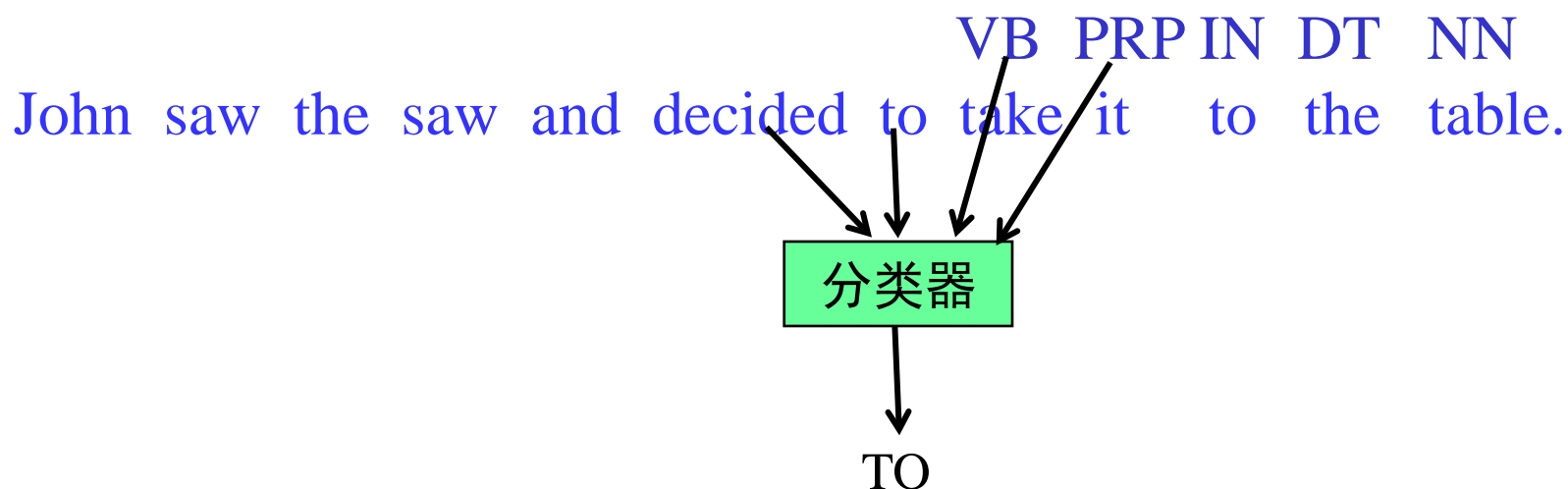
后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势



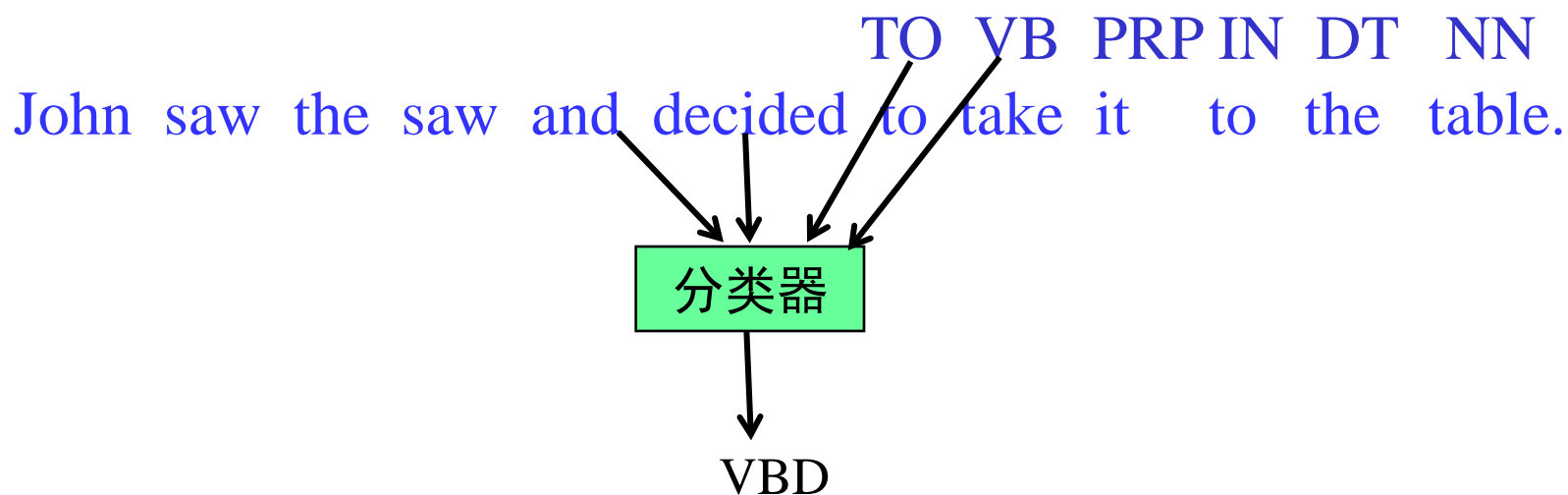
后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势



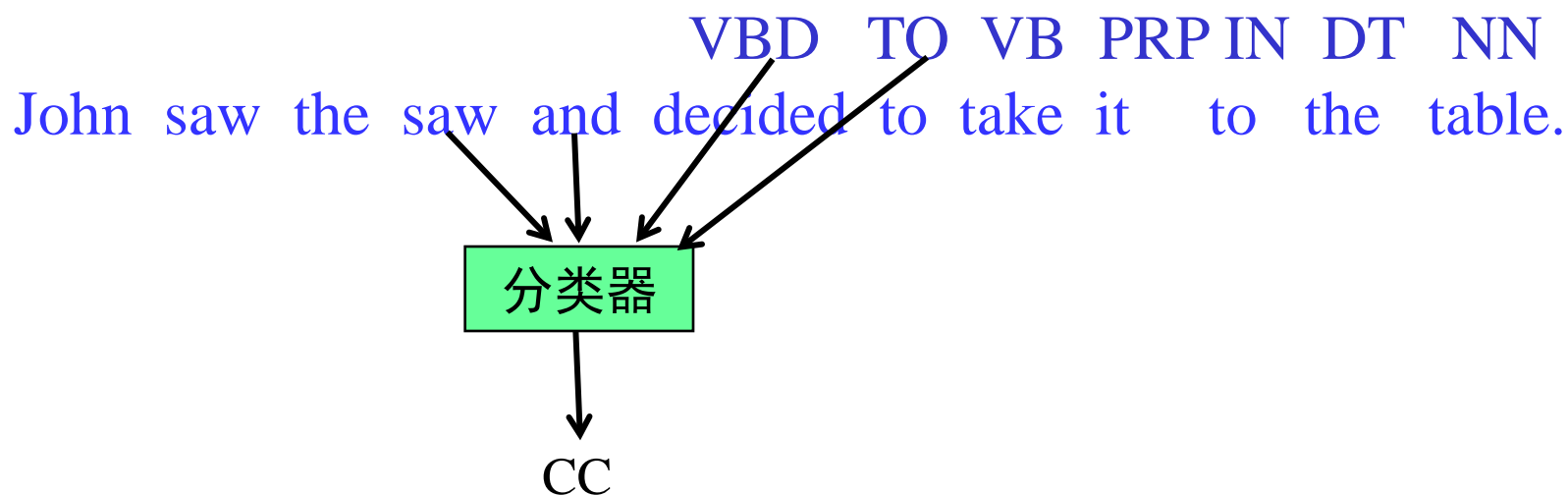
后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势



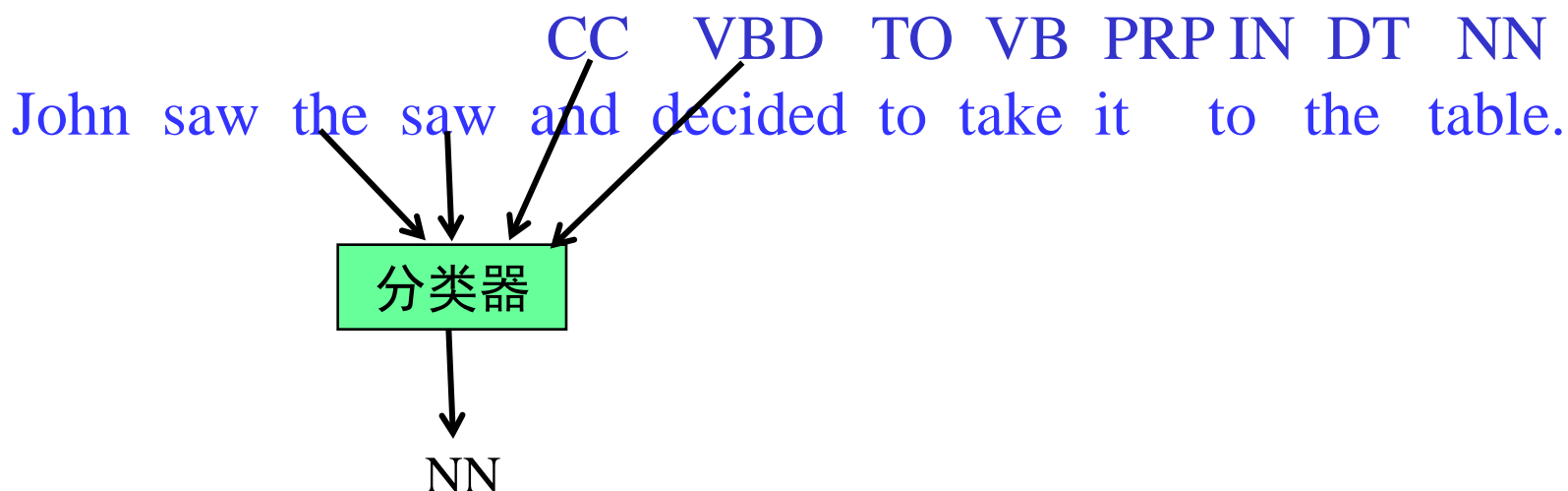
后向分类(Backward Classification)

- 对“to”进行分类的时候，后向算法比前向算法有优势



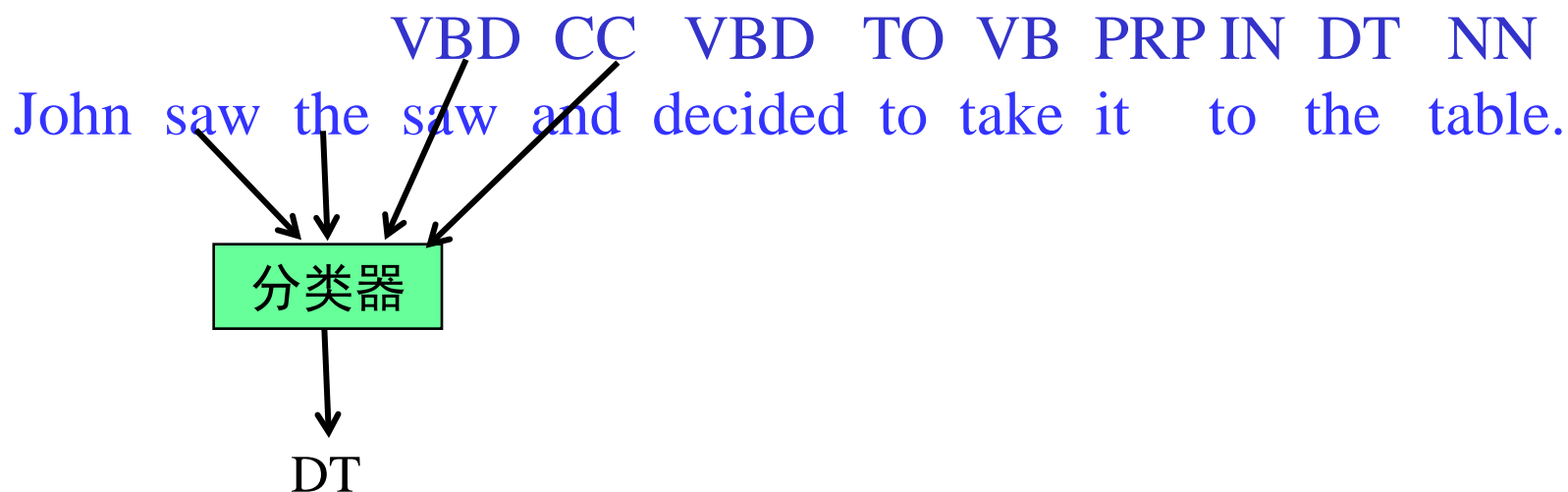
后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势



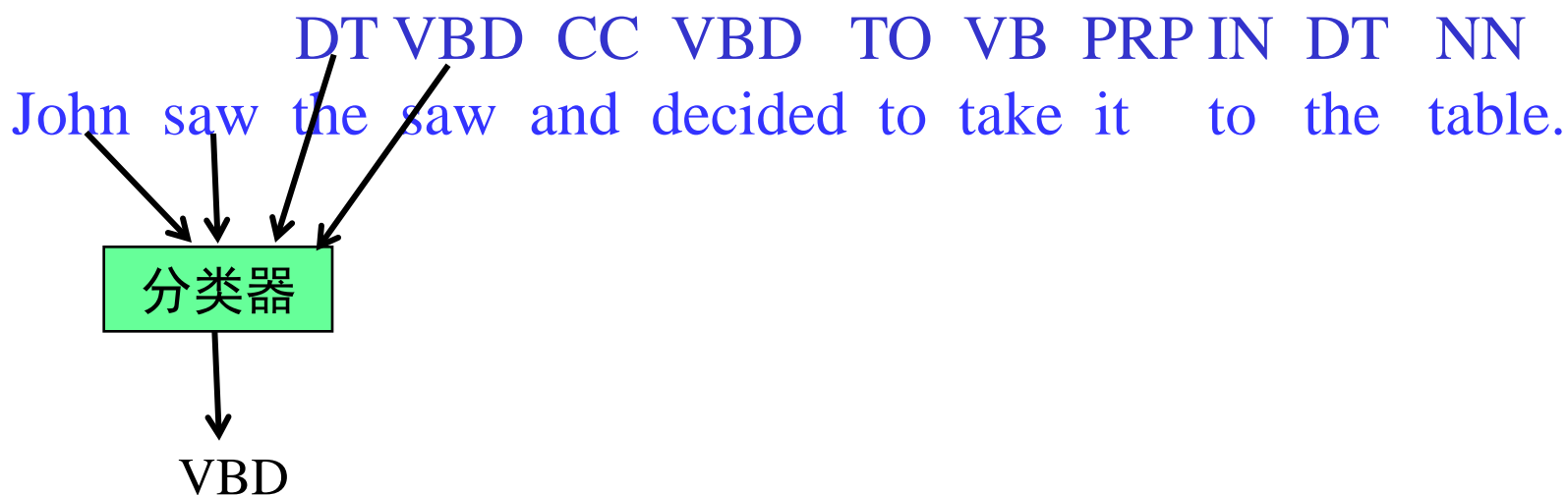
后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势



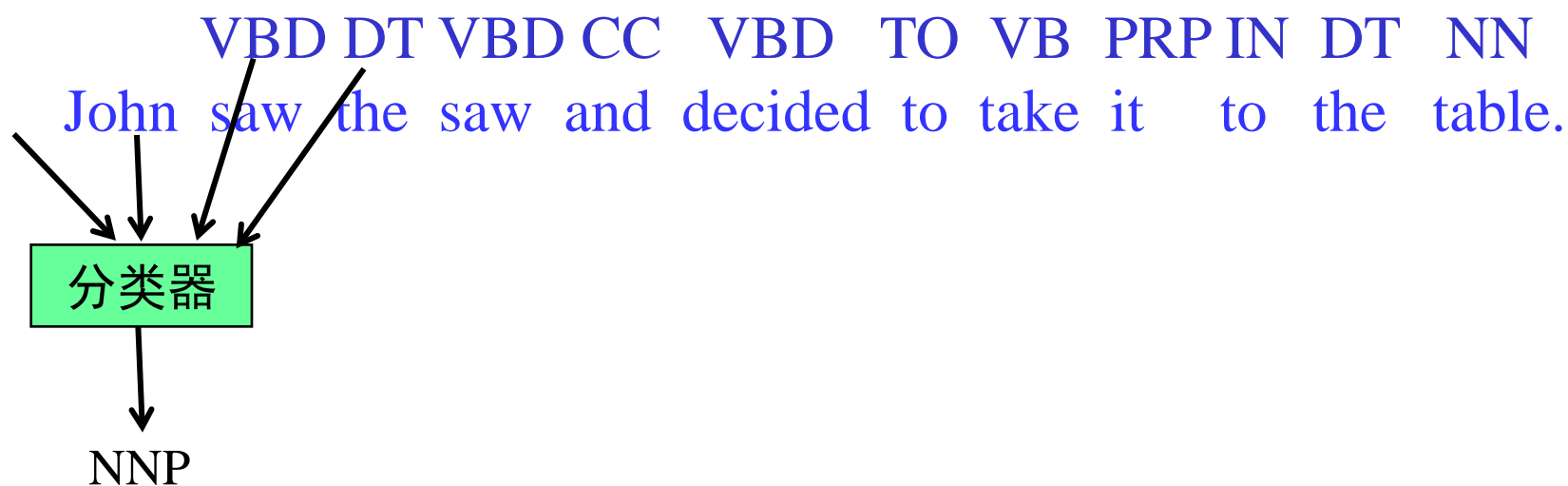
后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势



后向分类(Backward Classification)

- 对 “to” 进行分类的时候，后向算法比前向算法有优势

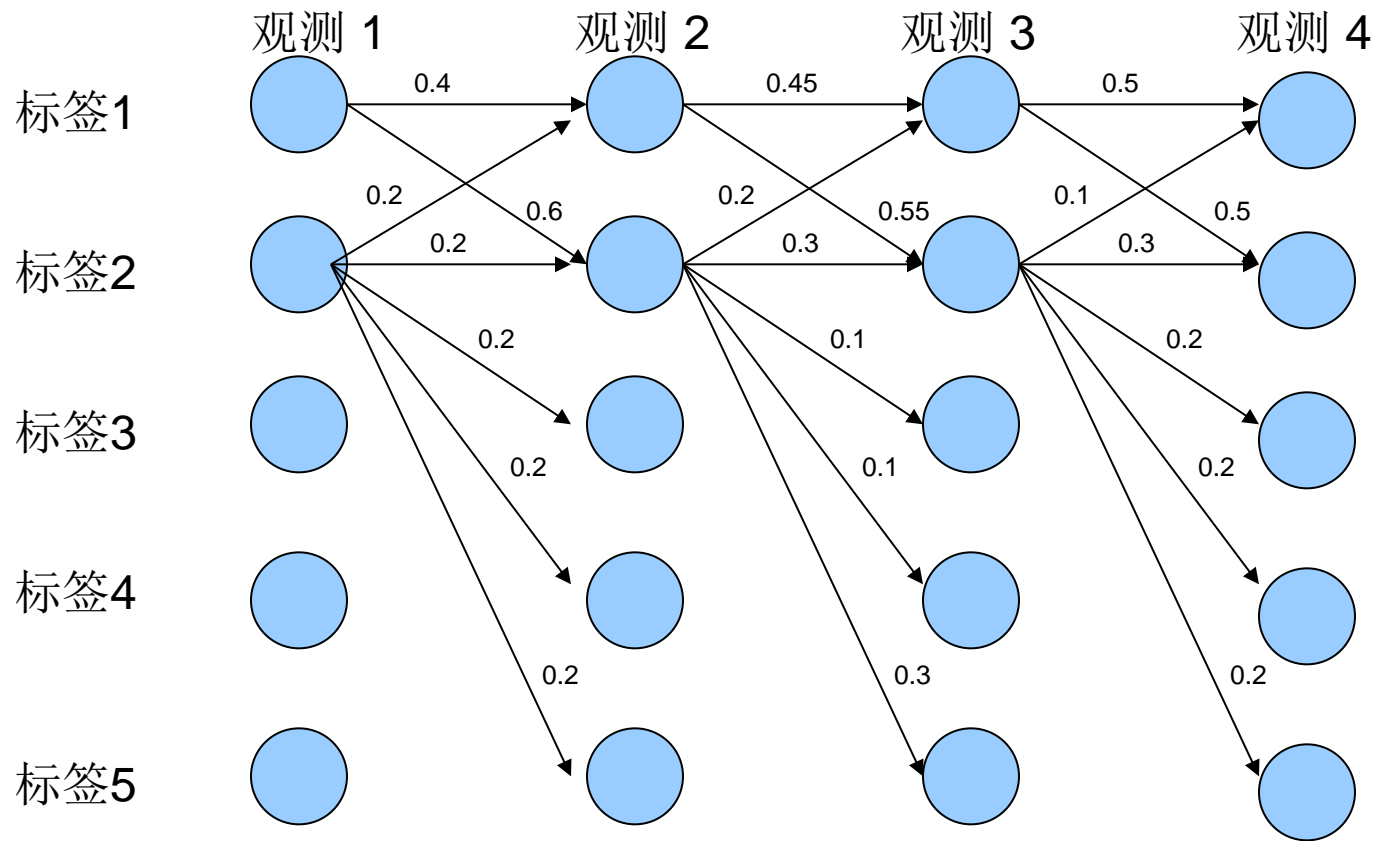


改进方法的问题

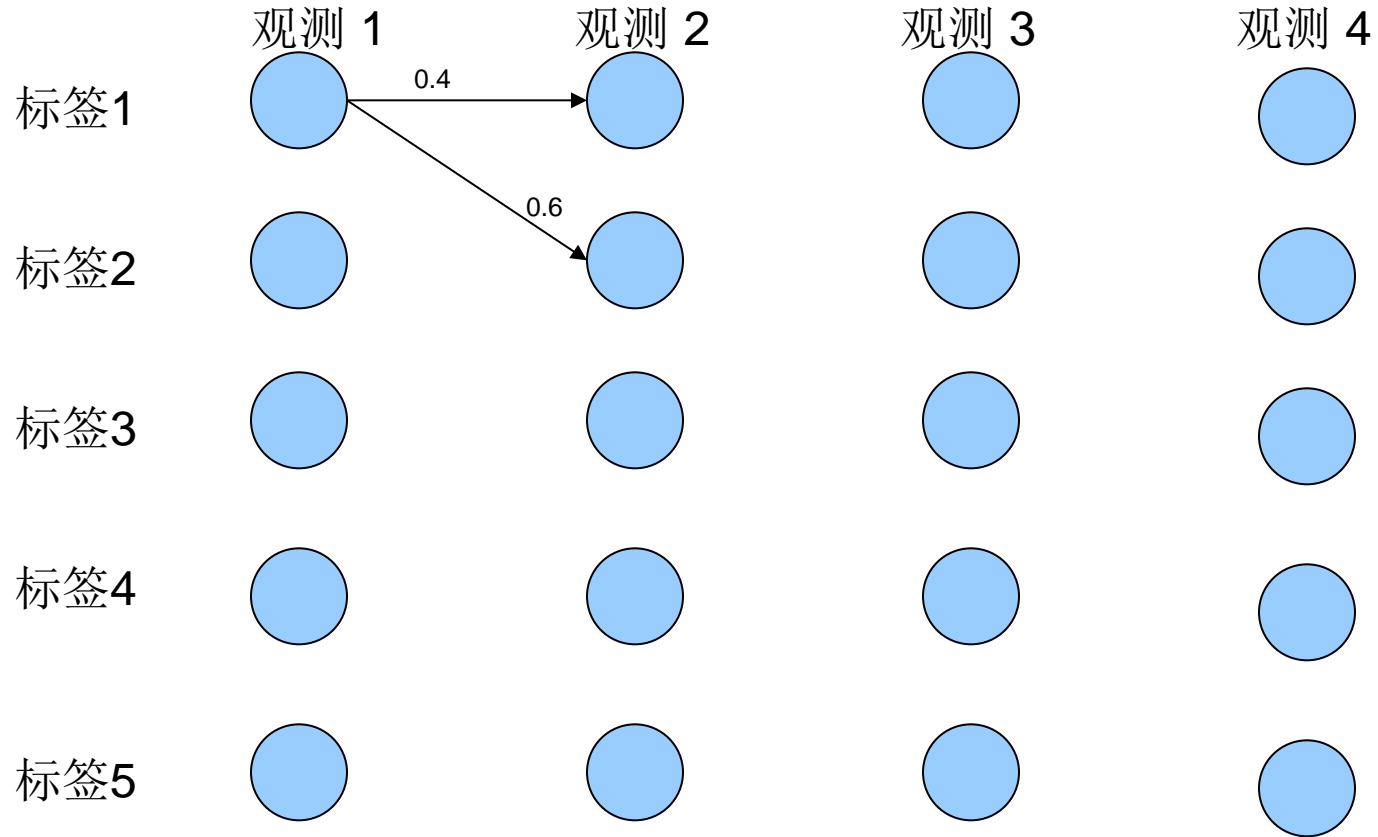
- 难以同时兼顾前向、后向的标签信息
- 每个决策仍然是局部最优的，难以统筹兼顾得到全局最优的决策
 - 全局最优的决策是指“同时”决定整个序列的标签
 - 局部最优的决策看似不错，但其实会有标注偏置问题 (label bias problem)



标注偏置问题(label bias problem)



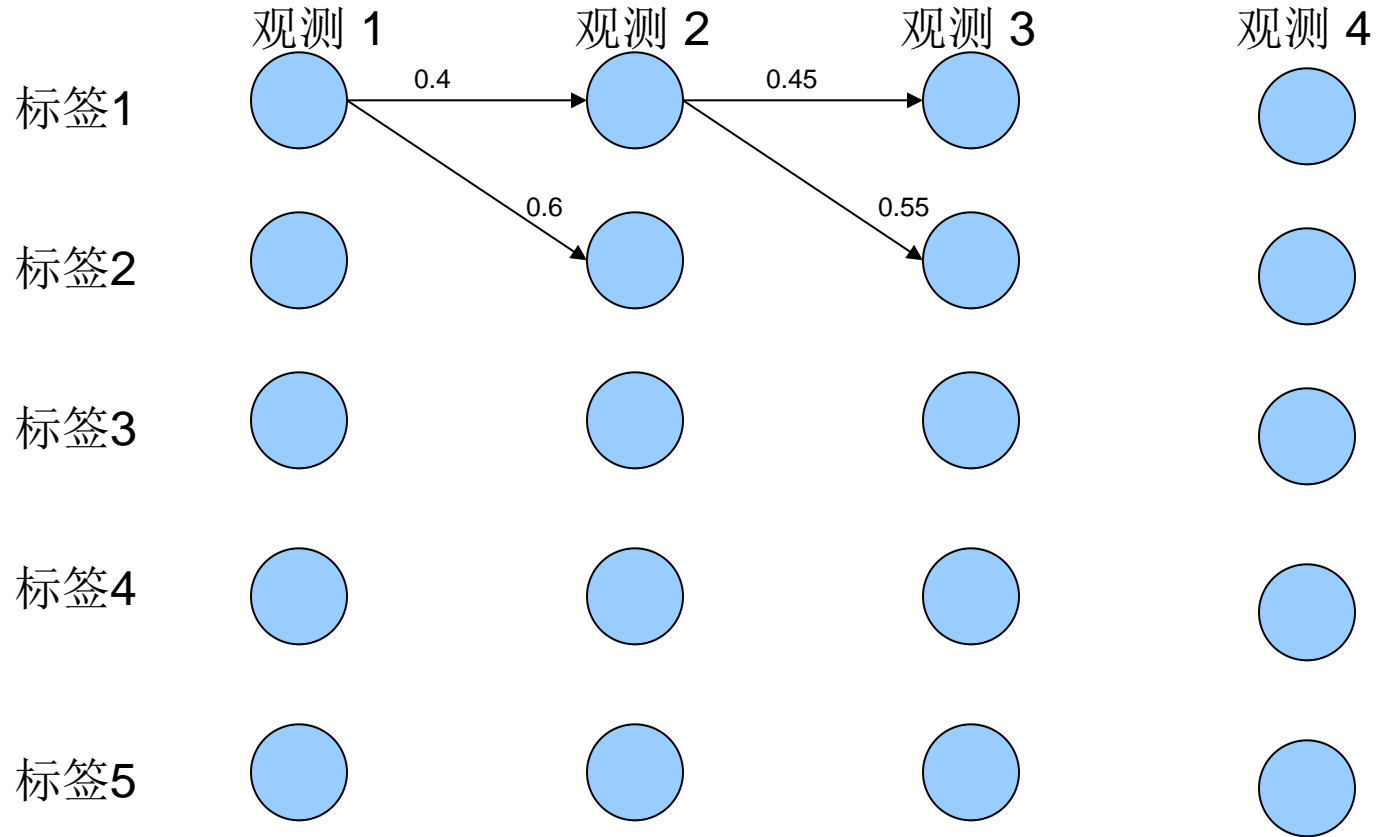
标注偏置问题(label bias problem)



从局部的概率转移情况可以看到:

- 标签1总是倾向于转移到标签2

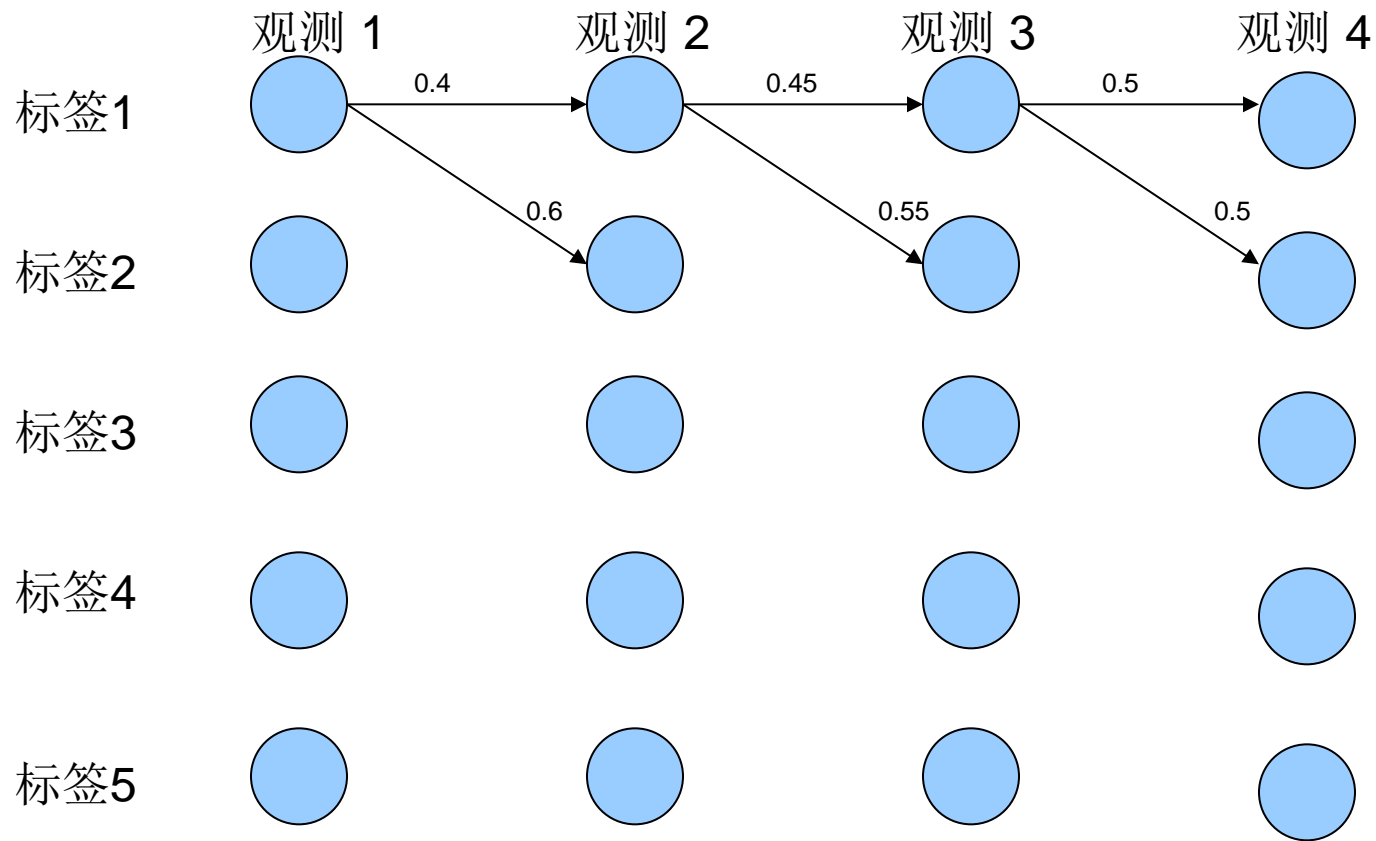
标注偏置问题(label bias problem)



从局部的概率转移情况可以看到:

- 标签1总是倾向于转移到标签2

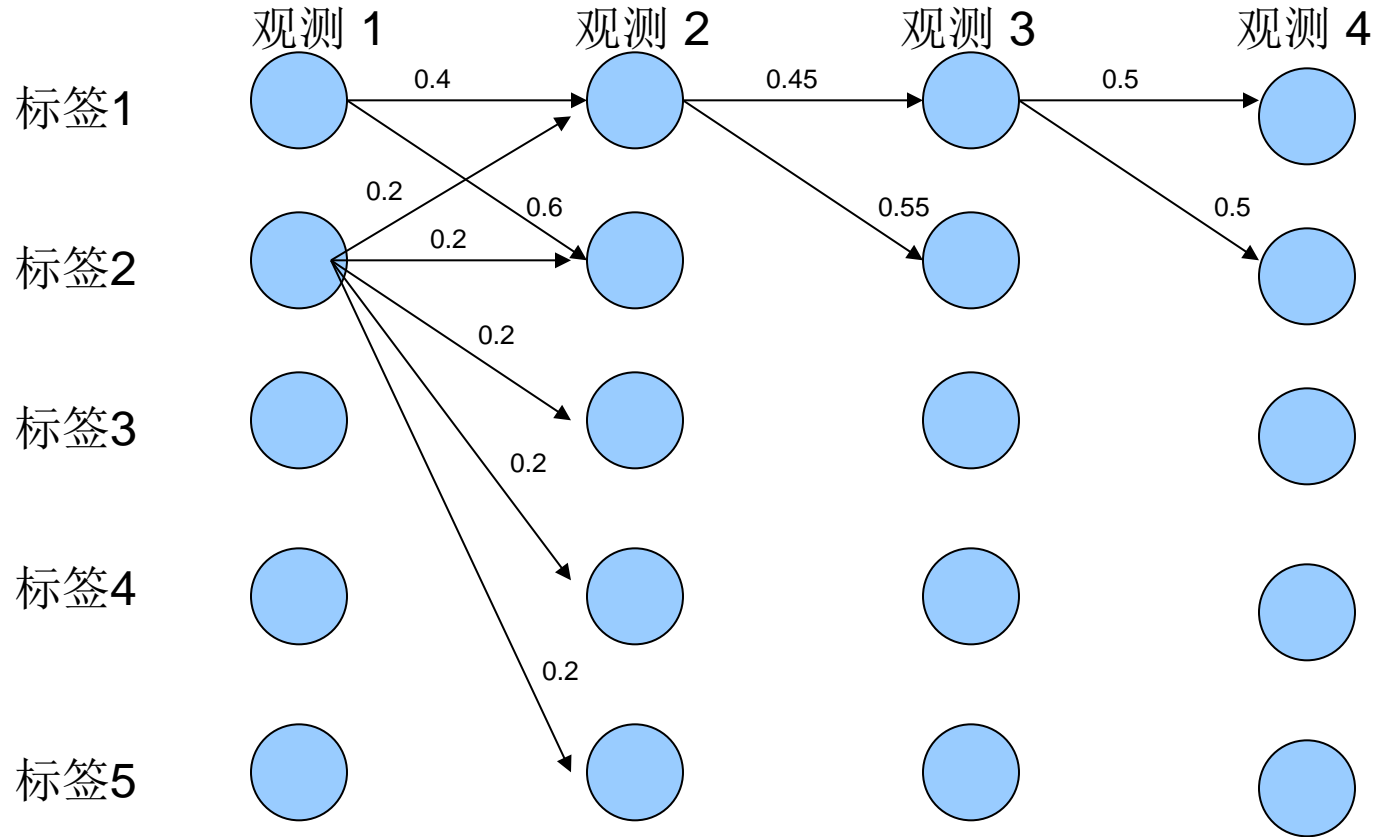
标注偏置问题(label bias problem)



从局部的概率转移情况可以看到:

- 标签1总是倾向于转移到标签2

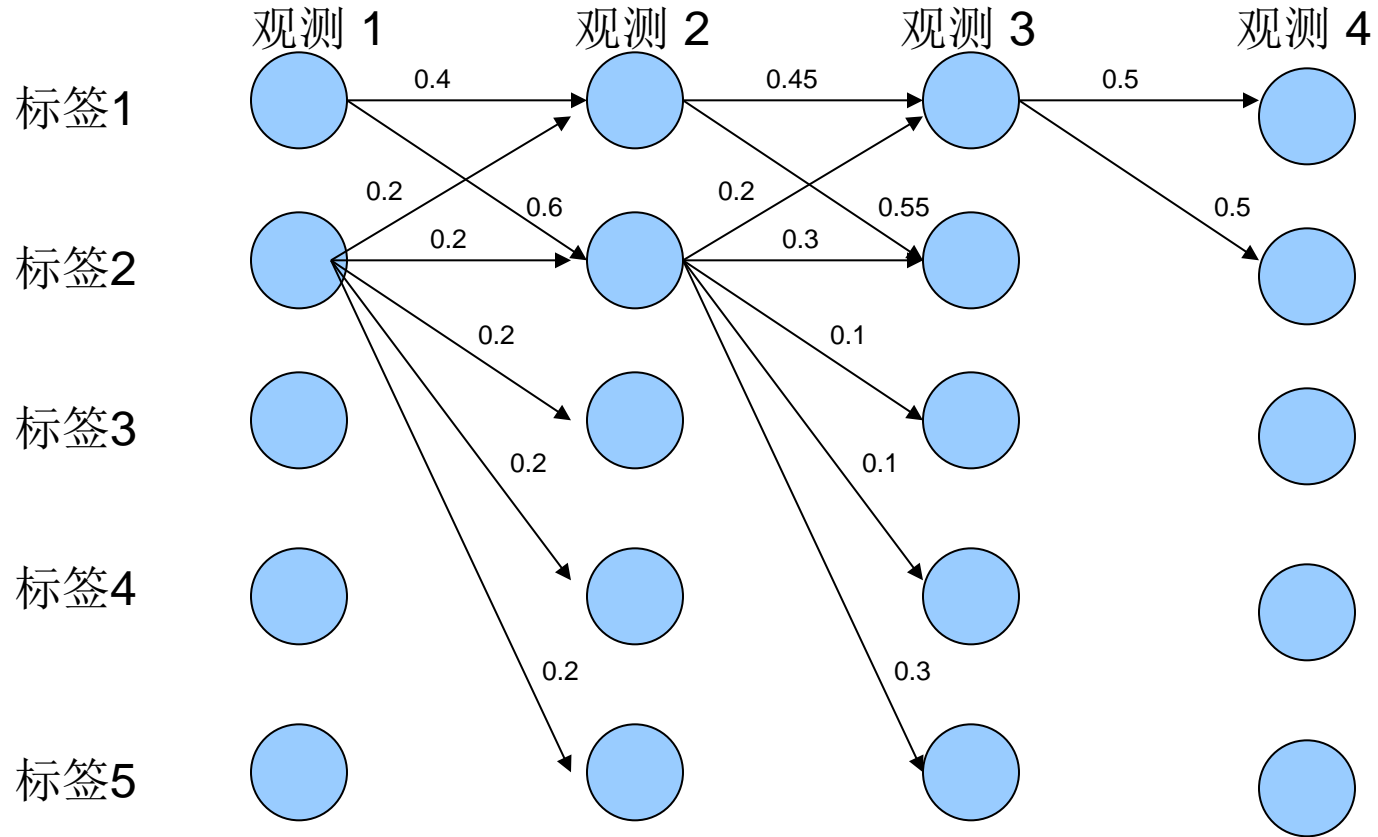
标注偏置问题(label bias problem)



从局部的概率转移情况可以看到:

- 标签1总是倾向于转移到标签2
- 标签2总是倾向于转移到标签2

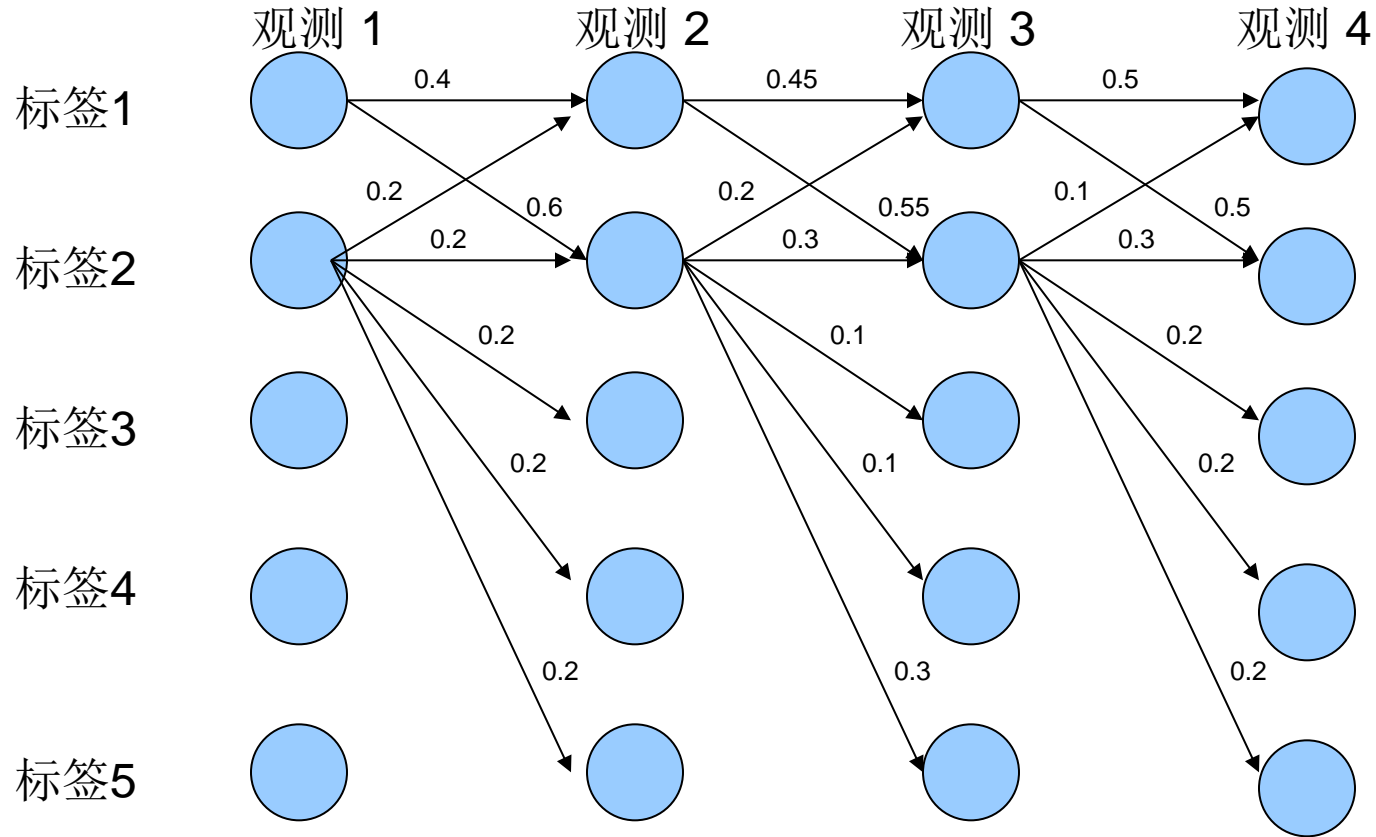
标注偏置问题(label bias problem)



从局部的概率转移情况可以看到:

- 标签1总是倾向于转移到标签2
- 标签2总是倾向于转移到标签2

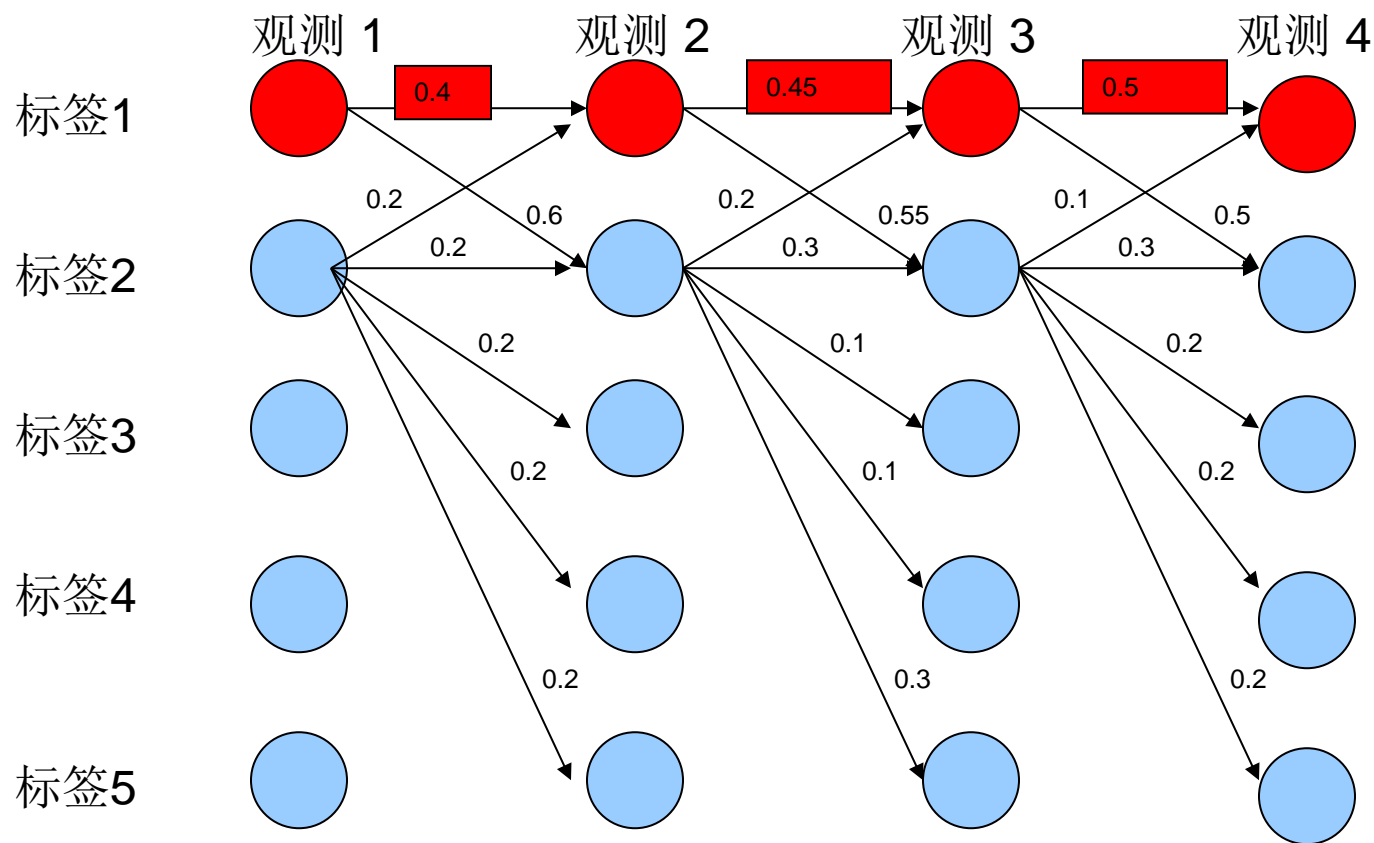
标注偏置问题(label bias problem)



从局部的概率转移情况可以看到:

- 标签1总是倾向于转移到标签2
- 标签2总是倾向于转移到标签2

标注偏置问题(label bias problem)



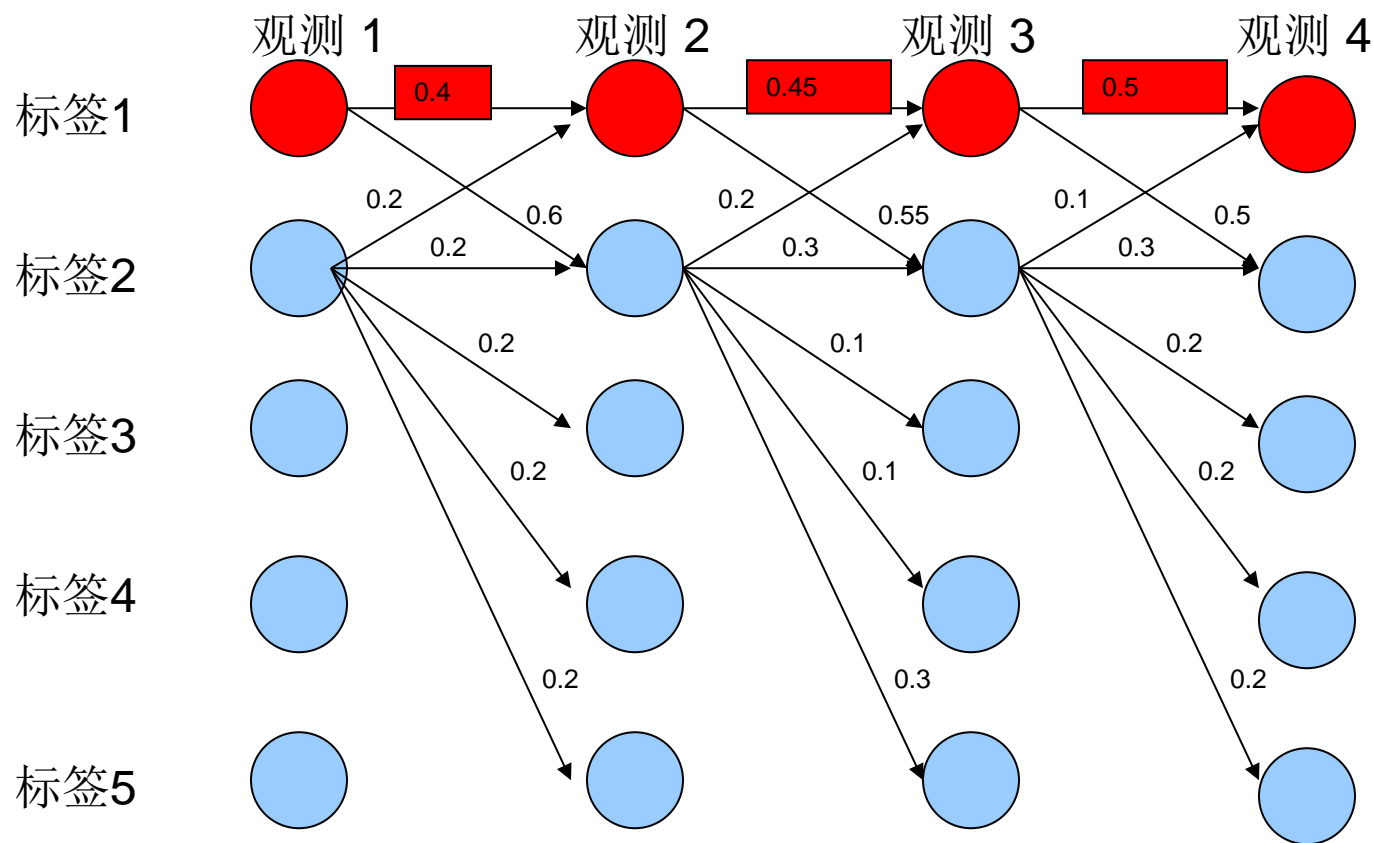
• 根据局部的概率转移情况，可得到局部最优的标签序列：

1 → 2 → 2 → 2 ($0.6 \times 0.3 \times 0.3 = 0.054$)

• 而实际上，全局最优的标签序列是：

1 → 1 → 1 → 1 ($0.4 \times 0.45 \times 0.5 = 0.09$)

标注偏置问题(label bias problem)



• 根据局部的概率转移情况，**得不到局部最优的标签序列**：

1 → 2 → 2 → 2 ($0.6 \times 0.3 \times 0.3 = 0.054$)

• 而实际上，**全局最优的标签序列是**：

1 → 1 → 1 → 1 ($0.4 \times 0.45 \times 0.5 = 0.09$)

(全局)序列标注模型

□ (全局)序列标注模型

- 能够对序列标注问题进行一个**全局的建模**，并确定一个**全局最优的决策**。从而解决局部最优导致的问题，例如标注偏置问题

常用的(全局)序列标注模型

- 隐马尔可夫模型 Hidden Markov Model (HMM)
- 结构化感知器 Structured Perceptron

□ 链状结构即通常所说的“序列标注问题”

□ 自然语言处理的序列标注问题举例

- 词性标注
- 中文切词
- 短语识别（浅层句法分析）
- 命名实体识别

□ 代表性的序列标注方法

- 关键问题是什么？
- 隐马尔可夫模型 HMM
- 结构化感知器 structured perceptron

开始讲解具体的序列标注方法

□ 参考书

- 《统计自然语言处理》第6章：
 - 概率图模型
 - Page 104 - 127

- 《统计自然语言处理》第7章：
 - 自动分词、命名实体识别与词性标注
 - Page 129 - 177