

Введение в обработку ТЕКСТОВ

Лекция 6

Статистические методы синтаксического анализа

Мотивация

- СКС-грамматики позволяют определить лучшее дерево разбора (т.е. устранить многозначность)
- Более точное моделирование языка, по сравнению с n-граммами
 - распознавание речи
 - машинный перевод
 - извлечение информации
 - ...
 - выделение ключевых слов

План

- Стохастические контекстно-свободные грамматики (СКС)
 - разрешение синтаксической многозначности
 - моделирование языка
- Вероятностная версия алгоритма СКУ
- Обучение СКС
- Проблемы СКС
 - разделение и слияние нетерминалов
 - СКС с поддержкой лексики
 - алгоритм Коллинза
- Методы оценки

Стохастические контекстно-свободные грамматики

N	множество нетерминальных символов
Σ	множество терминальных символов (непересекающееся с N)
R	множество правил, каждое вида $A \rightarrow \beta[p]$ где A - нетерминал, β - строка символов из множества $(\Sigma \cup N)^*$ p - вероятность правила $P(\beta A)$, $\sum_{\beta} P(A \rightarrow \beta) = 1$
S	символ начала

Пример

Грамматика	Вероятность	Лексикон
S → NP VP	0.8	Det → the a that this
S → Aux NP VP	0.1	0.6 0.2 0.1 0.1
S → VP	0.1	Noun → book flight meal money
NP → Pronoun	0.2	0.1 0.5 0.2 0.2
NP → Proper-Noun	0.2	Verb → book include prefer
NP → Det Nominal	0.6	0.5 0.2 0.3
Nominal → Noun	0.3	Pronoun → I he she me
Nominal → Nominal Noun	0.2	0.5 0.1 0.1 0.3
Nominal → Nominal PP	0.5	Proper-Noun → Houston NWA
VP → Verb	0.2	0.8 0.2
VP → Verb NP	0.5	Aux → does
VP → VP PP	0.3	1.0
PP → Prep NP	1.0	Prep → from to on near through
		0.25 0.25 0.1 0.2 0.2

Разрешение многозначности

- Вероятность разбора

$$P(T, S) = \prod_{i=1}^n P(RHS_i | LHS_i)$$

- Вероятность $P(T, S) = P(T)P(S|T) = P(T)$

- Выбор наиболее вероятного дерева

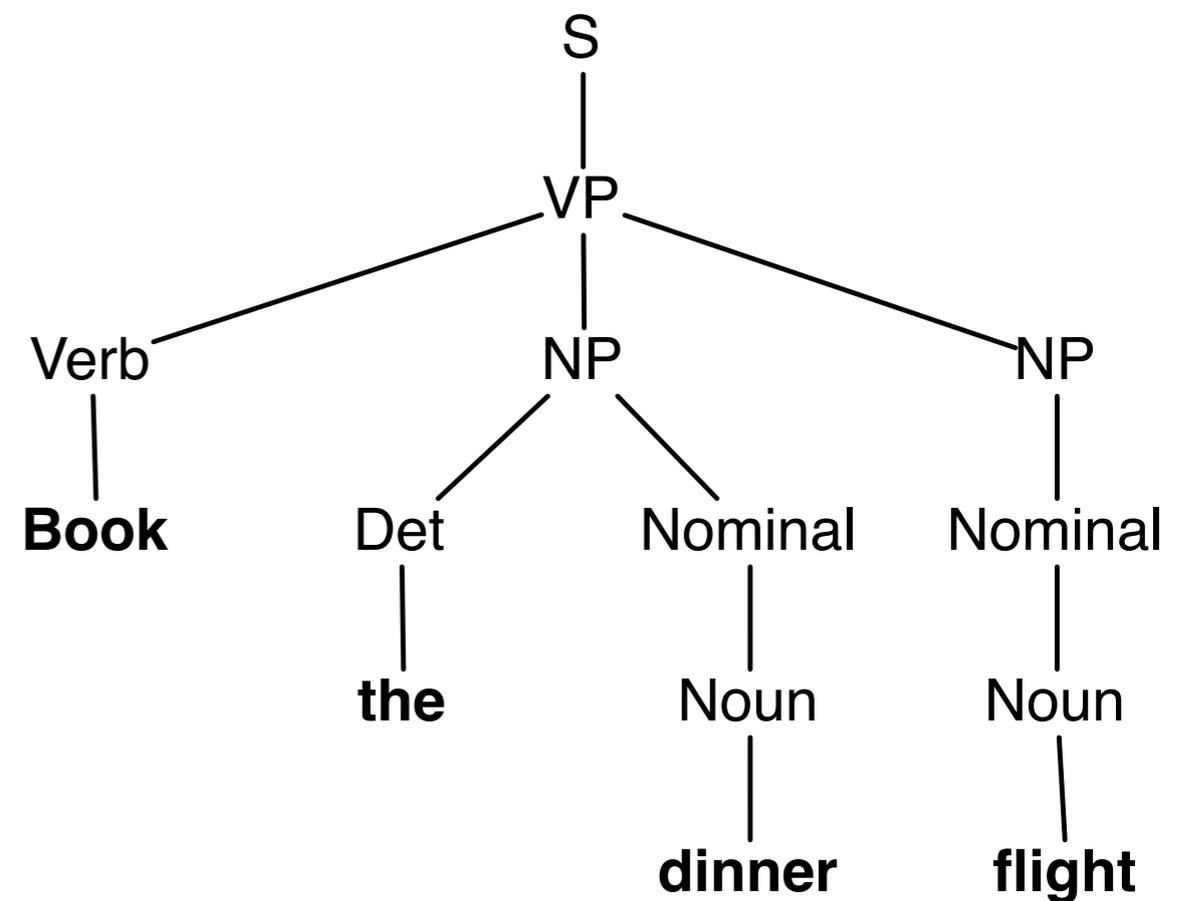
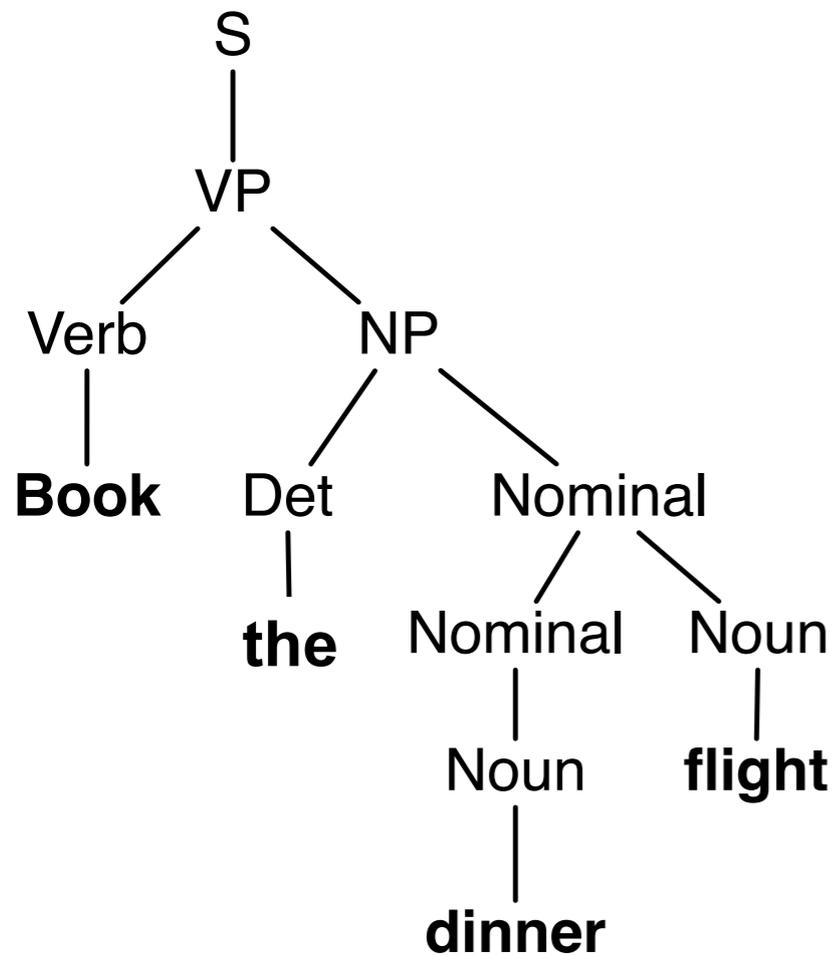
разбора $\hat{T}(S) = \arg \max_T P(T|S)$

$$\hat{T}(S) = \arg \max_T \frac{P(T, S)}{P(S)}$$

$$\hat{T}(S) = \arg \max_T P(T, S)$$

$$\hat{T}(S) = \arg \max_T P(T)$$

Разрешение многозначности



$$P(\text{T-left}) = .05 \cdot .20 \cdot .20 \cdot .20 \cdot .75 \cdot .30 \cdot .60 \cdot .10 \cdot .40 = 2.2 \cdot 10^{-6}$$

$$P(\text{T-right}) = .05 \cdot .10 \cdot .20 \cdot .15 \cdot .75 \cdot .75 \cdot .30 \cdot .60 \cdot .10 \cdot .40 = 6.1 \cdot 10^{-7}$$

Моделирование языка

$$P(S) = \sum_T P(T, S) = \sum_T P(T)$$

- Вариант 1:
 - Этап 1: с помощью n-граммной модели получить m лучших предложений
 - Этап 2: выбрать наиболее вероятное предложение на основе грамматики
- Вариант 2:
 - Модифицировать парсер для предсказания следующего слова (Xu et. al 2002)

Вероятностная версия алгоритма СКУ

- Добавляем в каждую ячейку вероятность нетерминального символа
- Ячейка $[i, j]$ должна содержать наиболее вероятный вывод, покрывающий с $i+1$ по j слова и содержать их вероятность.
- При трансформации грамматики к нормальной форме необходимо сохранить вероятности правил

Преобразование грамматики

Оригинальная грамматика

Грамматика в нормальной форме Хомского

$S \rightarrow NP VP$	0.8	$S \rightarrow NP VP$	0.8
$S \rightarrow Aux NP VP$	0.1	$S \rightarrow X1 VP$	0.1
		$X1 \rightarrow Aux NP$	1.0
$S \rightarrow VP$	0.1	$S \rightarrow \text{book} \mid \text{include} \mid \text{prefer}$	
		0.01 0.004 0.006	
		$S \rightarrow \text{Verb NP}$	0.05
		$S \rightarrow VP PP$	0.03
$NP \rightarrow \text{Pronoun}$	0.2	$NP \rightarrow \text{I} \mid \text{he} \mid \text{she} \mid \text{me}$	
		0.1 0.02 0.02 0.06	
$NP \rightarrow \text{Proper-Noun}$	0.2	$NP \rightarrow \text{Houston} \mid \text{NWA}$	
		0.16 .04	
$NP \rightarrow \text{Det Nominal}$	0.6	$NP \rightarrow \text{Det Nominal}$	0.6
$\text{Nominal} \rightarrow \text{Noun}$	0.3	$\text{Nominal} \rightarrow \text{book} \mid \text{flight} \mid \text{meal} \mid \text{money}$	
		0.03 0.15 0.06 0.06	
$\text{Nominal} \rightarrow \text{Nominal Noun}$	0.2	$\text{Nominal} \rightarrow \text{Nominal Noun}$	0.2
$\text{Nominal} \rightarrow \text{Nominal PP}$	0.5	$\text{Nominal} \rightarrow \text{Nominal PP}$	0.5
$VP \rightarrow \text{Verb}$	0.2	$VP \rightarrow \text{book} \mid \text{include} \mid \text{prefer}$	
		0.1 0.04 0.06	
$VP \rightarrow \text{Verb NP}$	0.5	$VP \rightarrow \text{Verb NP}$	0.5
$VP \rightarrow VP PP$	0.3	$VP \rightarrow VP PP$	0.3
$PP \rightarrow \text{Prep NP}$	1.0	$PP \rightarrow \text{Prep NP}$	1.0

Вероятностная версия алгоритма SKY

Book the flight through Houston

S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None			
	Det:6 ←	NP:.6*.6*.15 =.054		
		Nominal:.15 Noun:.5		

Вероятностная версия алгоритма SKY

Book the flight through Houston

S :.01, VP:.1, Verb:.5 ← Nominal:.03 Noun:.1	None	$VP:.5 \cdot .5 \cdot .054 = .0135$		
	Det:.6	$NP:.6 \cdot .6 \cdot .15 = .054$		
		Nominal:.15 Noun:.5		

Вероятностная версия алгоритма SKY

Book the flight through Houston

S :.01, VP:.1, Verb:.5 ← Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135		
	Det:.6	NP:.6*.6*.15 =.054		
		Nominal:.15 Noun:.5		

Вероятностная версия алгоритма SKY

Book the flight through Houston

S :.01, VP:~.1, Verb:~.5 Nominal:~.03 Noun:~.1	None	S:~.05*~.5*~.054 =.00135 VP:~.5*~.5*~.054 =.0135	None	
	Det:~.6	NP:~.6*~.6*~.15 =.054	None	
		Nominal:~.15 Noun:~.5	None	
			Prep:~.2	

Вероятностная версия алгоритма SKY

Book the flight through Houston

S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	
	Det:.6	NP:.6*.6*.15 =.054	None	
		Nominal:.15 Noun:.5	None	
			Prep:.2 ←	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

Вероятностная версия алгоритма SKY

Book the flight through Houston

S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	
	Det:.6	NP:.6*.6*.15 =.054	None	
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

Вероятностная версия алгоритма SKY

Book the flight through Houston

S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	
	Det:.6 ←	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

Вероятностная версия алгоритма SKY

Book the flight through Houston

S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1		S:.05*.5*.054 =.00135		S:.05*.5* .000864 =.0000216
	None	VP:.5*.5*.054 =.0135	None	
	Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

Вероятностная версия алгоритма SKY

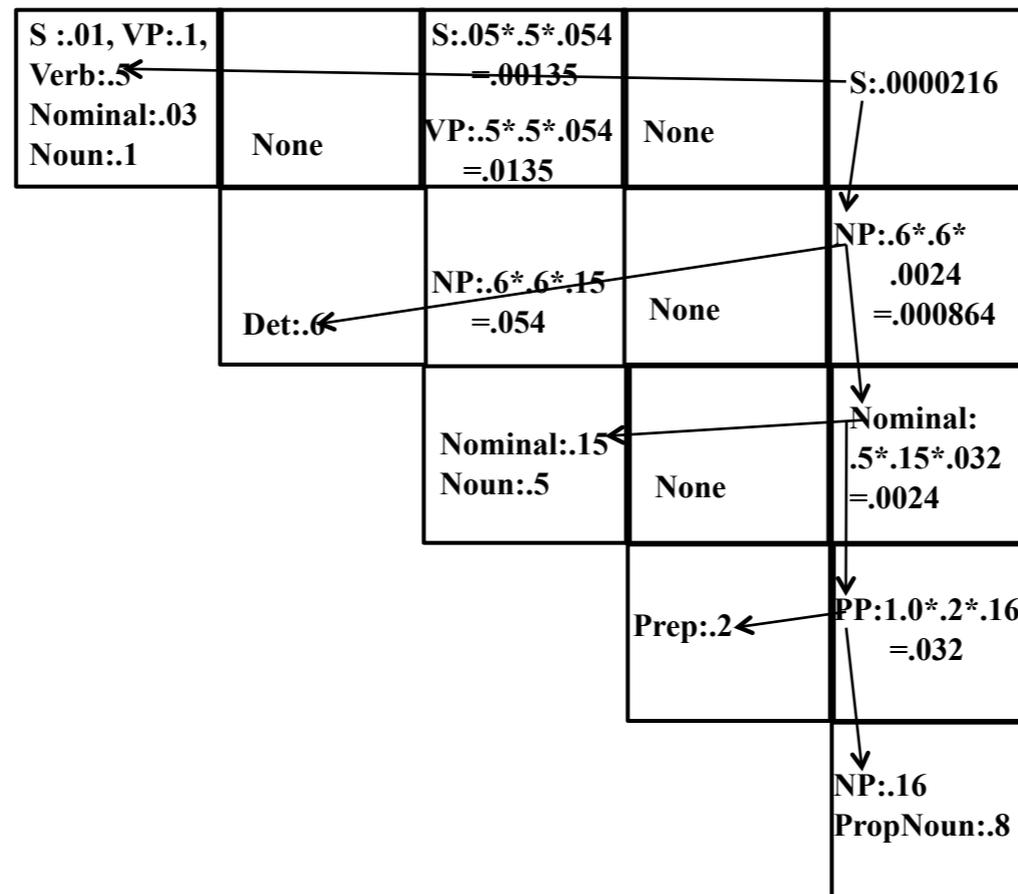
Book the flight through Houston

S :.01, VP:.1, Verb:.5 Nominal:.03 Noun:.1	None	S:.05*.5*.054 =.00135 VP:.5*.5*.054 =.0135	None	S:.03*.0135* .032 =.00001296 S:.0000216
	Det:.6	NP:.6*.6*.15 =.054	None	NP:.6*.6* .0024 =.000864
		Nominal:.15 Noun:.5	None	Nominal: .5*.15*.032 =.0024
			Prep:.2	PP:1.0*.2*.16 =.032
				NP:.16 PropNoun:.8

Вероятностная версия алгоритма SKY

Выбираем наиболее вероятное дерево разбора

Book the flight through Houston



Обучение СКС

- Вычисление вероятности на основе банка деревьев

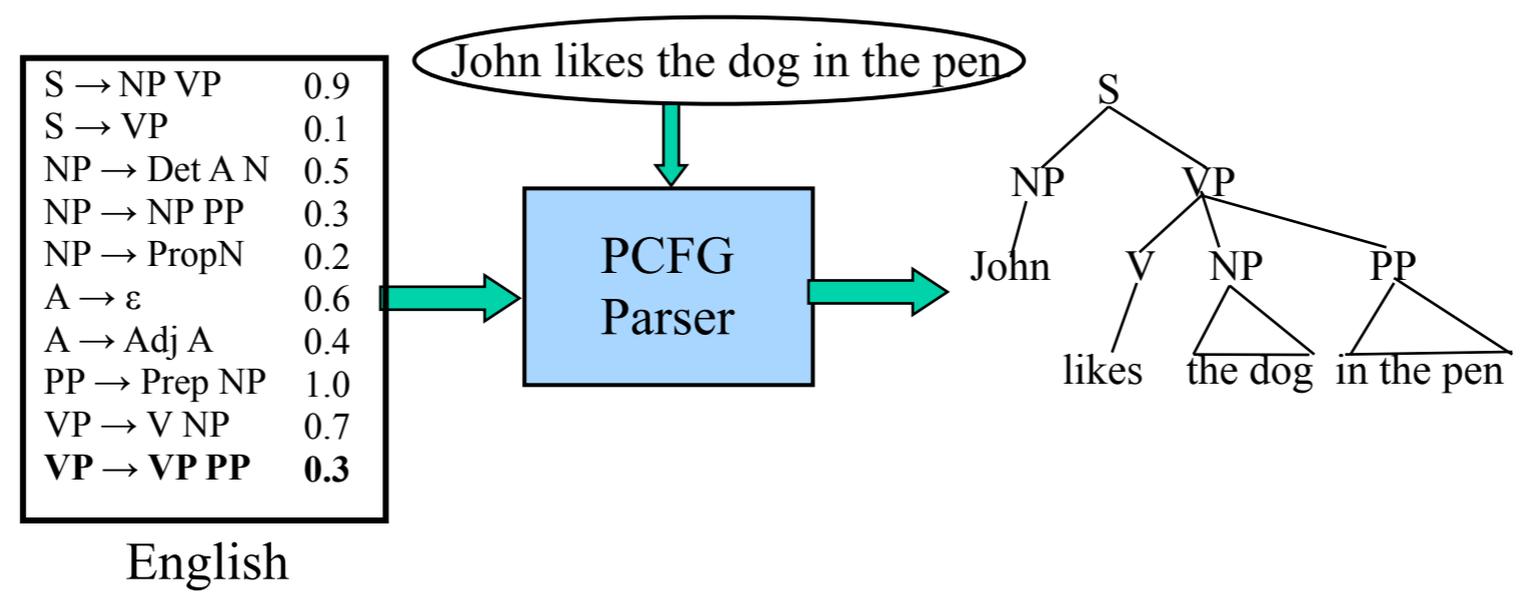
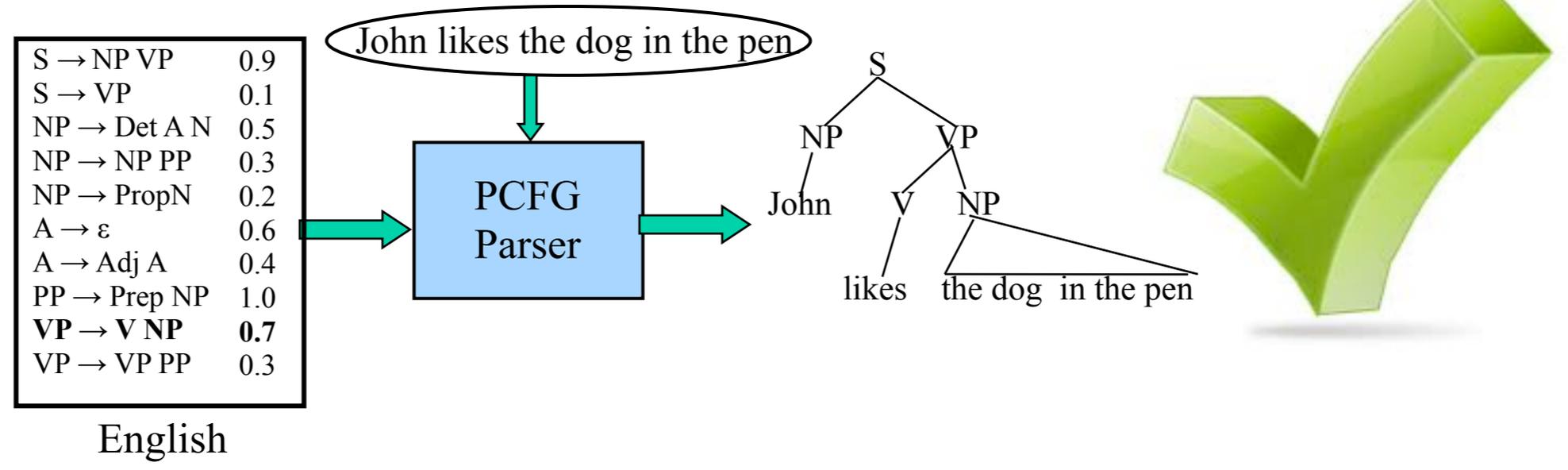
$$P(\alpha \rightarrow \beta | \alpha) = \frac{\text{Count}(\alpha \rightarrow \beta)}{\sum_{\gamma} \text{Count}(\alpha \rightarrow \gamma)} = \frac{\text{Count}(\alpha \rightarrow \beta)}{\text{Count}(\alpha)}$$

- Вывод без тренировочного множества (EM)
 - На основе множества предложений построить множество наиболее вероятных синтаксических разборов
 - Обновить значения вероятностей на основе полученных данных
 - (Manning and Schütze 1999)

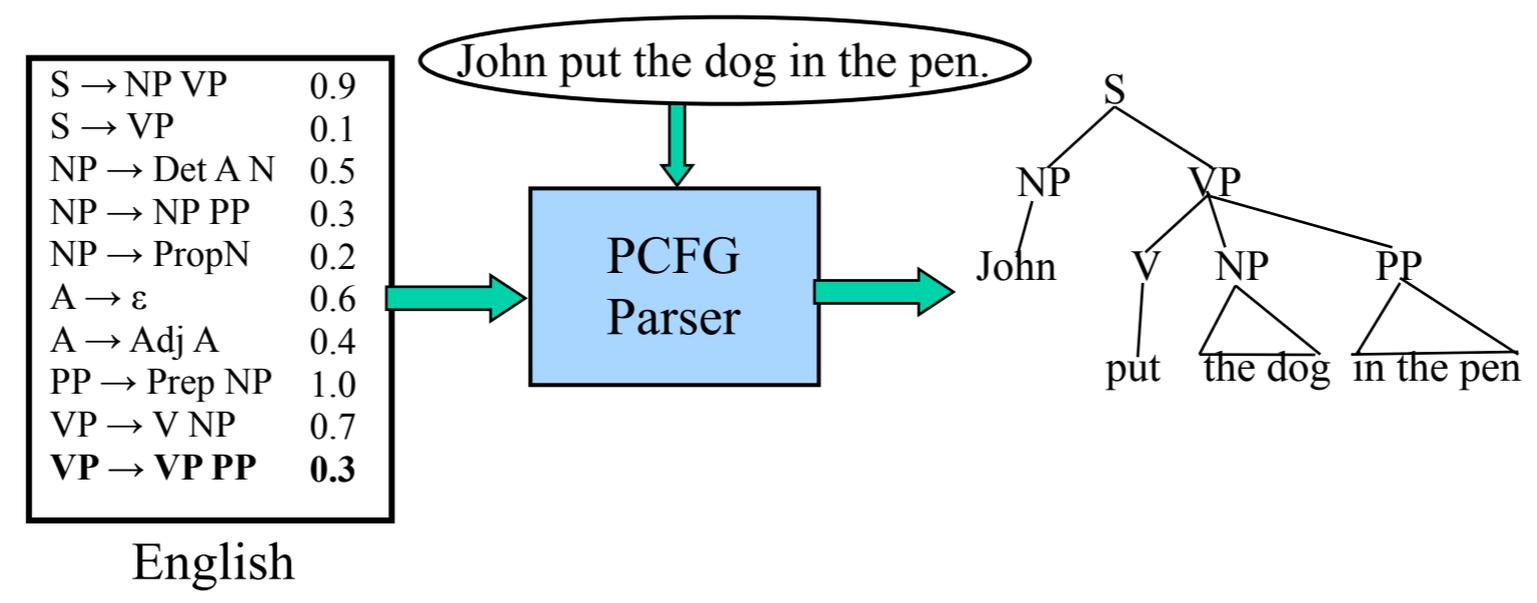
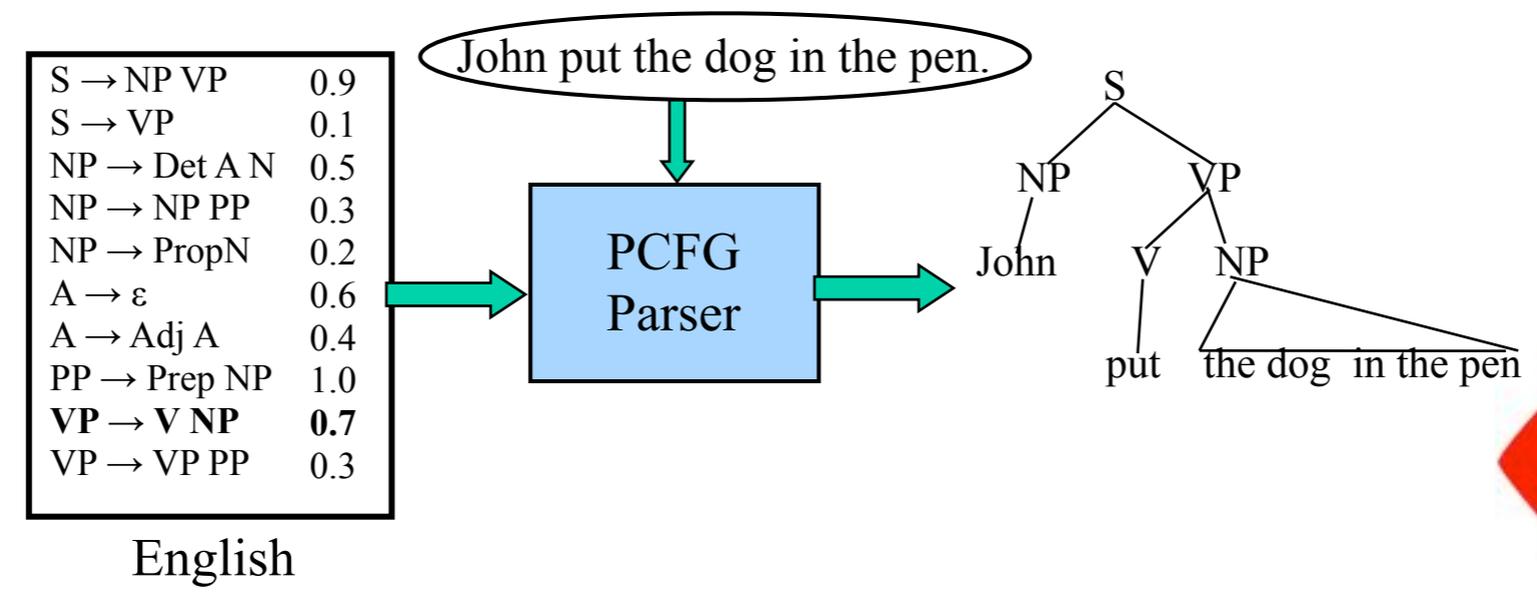
Проблемы СКС

- Предположение о независимости правил, не позволяет хорошо моделировать структурные зависимости в дереве разбора
- СКС не могут моделировать синтаксические факты о конкретных словах

Пример

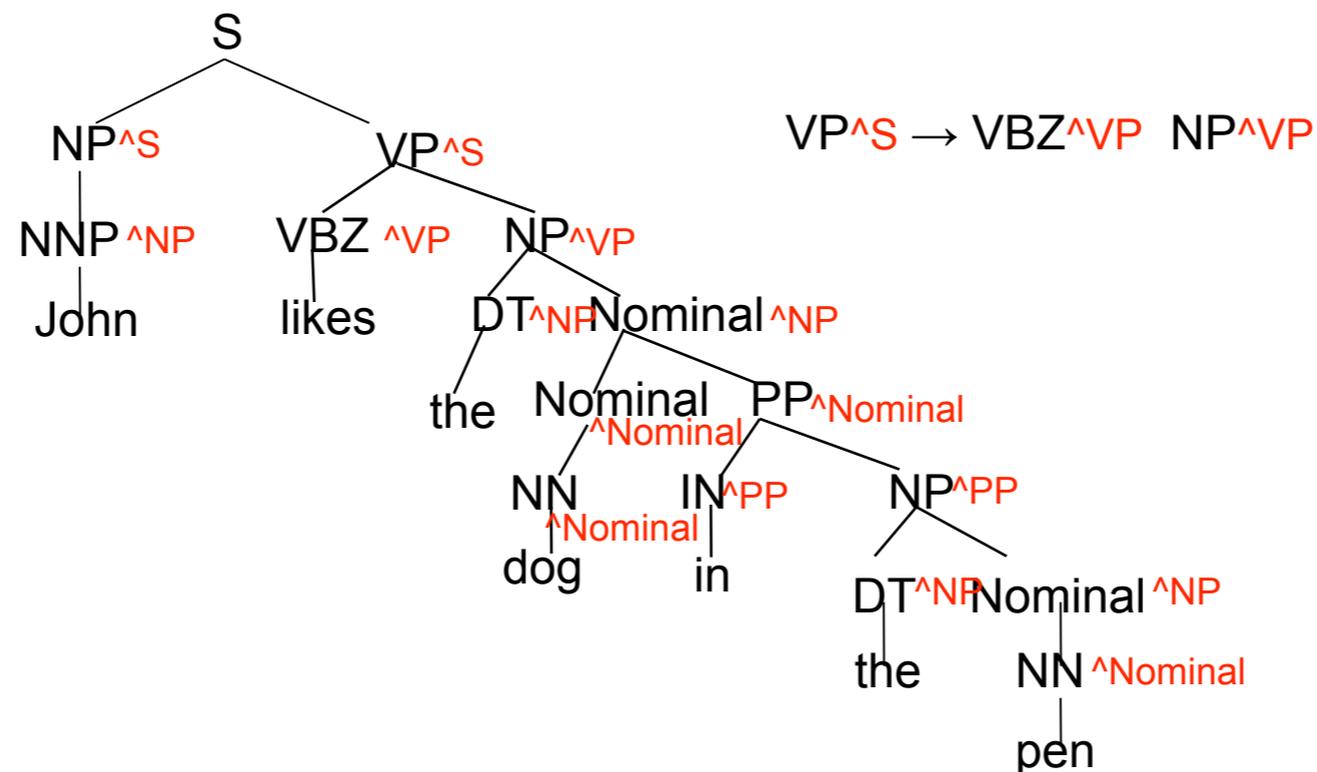


Пример



Решение проблемы зависимостей

- Для добавления контекстуальной информации нетерминалы можно разделить на несколько, используя родительские узлы в дереве разбора (parent annotation)

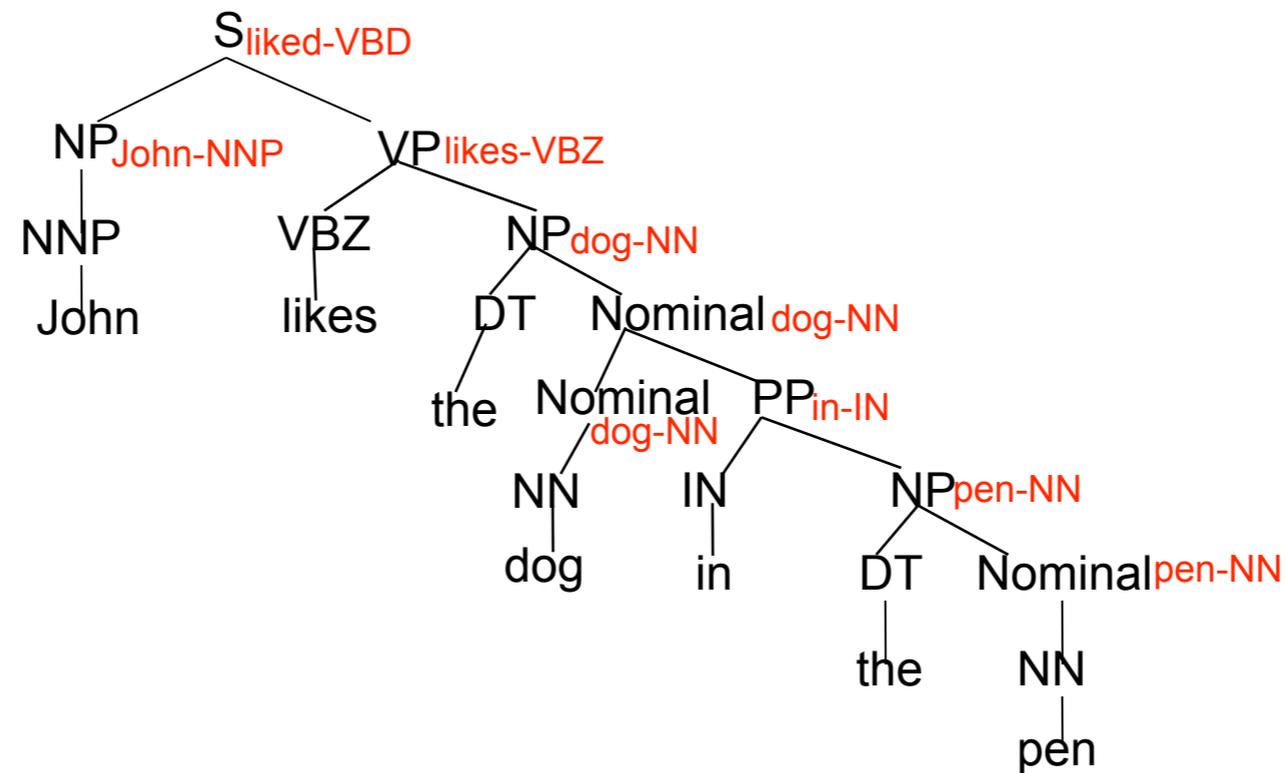


Разделение и слияние

- Разделение нетерминальных символов сильно увеличивает грамматику
- Лучше разделять нетерминалы, только если это приведет к улучшению точности
- Также можно объединять некоторые нетерминальные символы, чтобы достичь большей точности
- Метод: эвристический поиск наилучшей комбинации разделений и слияний которая будет максимизировать правдоподобие банка деревьев

СКС с поддержкой лексики

- Расширение правил
 - $VP \rightarrow VP PP$
 - $VP(\text{put}) \rightarrow VP(\text{put}) PP(\text{in})$
 - $VP(\text{put}, VDB) \rightarrow VP(\text{put}, VDB) PP(\text{in}, IN)$



Оценка вероятности

- Точная оценка невозможна, так как не существует достаточного количества деревьев
- Необходимо сделать еще предположения (о независимости), которые помогут оценить эти вероятности
- Алгоритмы (Collins, 1999), (Charniak, 1997)

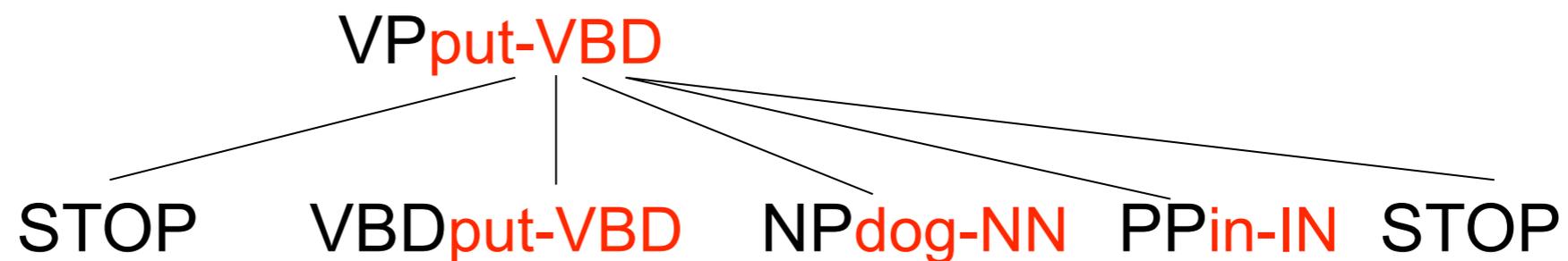
Алгоритм Коллинза

- Использует простую производящую модель
- $LHS \rightarrow L_n L_{n-1} \dots L_1 H R_1 \dots R_{m-1} R_m$
 - H-вершина группы
 - L - символы слева
 - R - символы справа
 - По краям символы STOP
- Вероятности левых и правых символов зависят только от вершины группы и нетерминала в левой части правила

Пример

$VP_{\text{put-VBD}} \rightarrow VBD_{\text{put-VBD}} NP_{\text{dog-NN}} PP_{\text{in-IN}}$

$VP_{\text{put-VBD}} \rightarrow$
 $STOP_{L_1}$
 $VBD_{\text{put-VBD}}_H$
 $NP_{\text{dog-NN}}_{R_1}$
 $PP_{\text{in-IN}}_{R_2}$
 $STOP_{R_3}$



$P(VP_{\text{put-VBD}} \rightarrow VBD_{\text{put-VBD}} NP_{\text{dog-NN}} PP_{\text{in-IN}}) =$

$= P_H(VBD_{\text{put-VBD}} | VP_{\text{put-VBD}}) *$

$* P_L(STOP | VBD_{\text{put-VBD}}, VP_{\text{put-VBD}}) *$

$* P_R(NP_{\text{dog-NN}} | VBD_{\text{put-VBD}}, VP_{\text{put-VBD}}) *$

$* P_R(PP_{\text{in-IN}} | VBD_{\text{put-VBD}}, VP_{\text{put-VBD}}) *$

$* P_R(STOP | VBD_{\text{put-VBD}}, VP_{\text{put-VBD}})$

Оценка вероятностей

- Вероятности можно оценить на основе банка деревьев

$$P_R(\text{PPin-IN} \mid \text{VPput-VBD}) = \frac{\text{Count}(\text{PPin-IN справа от вершины в правиле для VPput-VBD})}{\text{Count}(\text{символы справа от вершины в правиле для VPput-VBD})}$$

- Сглаживание можно осуществлять через откат или линейную интерполяцию

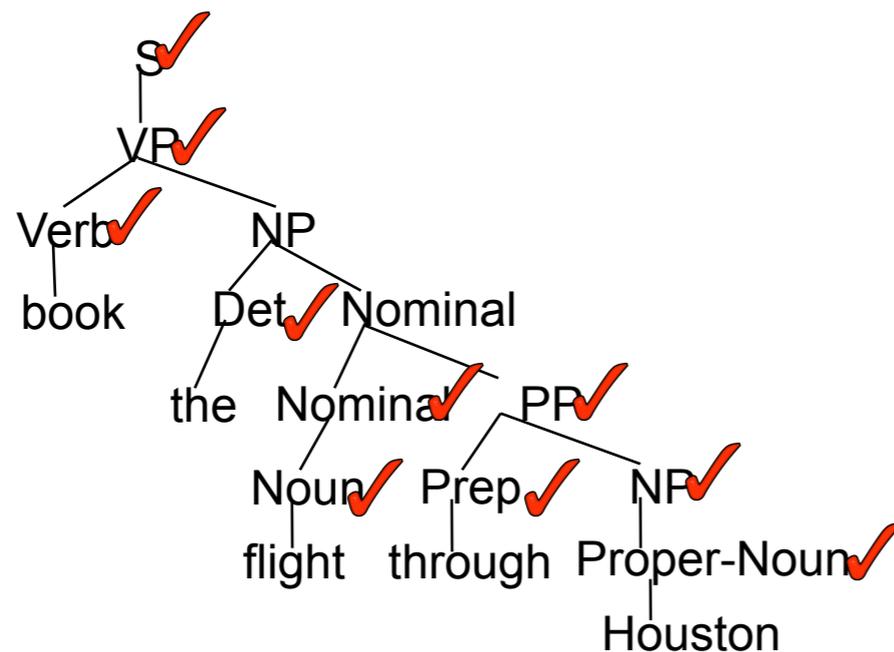
$$\begin{aligned} \text{sm}P_R(\text{PPin-IN} \mid \text{VPput-VBD}) &= \lambda_1 P_R(\text{PPin-IN} \mid \text{VPput-VBD}) \\ &+ (1 - \lambda_1) (\lambda_2 P_R(\text{PPin-IN} \mid \text{VPVBD}) + \\ &\quad (1 - \lambda_2) P_R(\text{PPin-IN} \mid \text{VP})) \end{aligned}$$

Оценка качества алгоритма

- Метрика PARSEVAL: пусть P - дерево разбора, созданное алгоритмом, T - дерево разбора, созданное экспертами
 - Точность = $(\# \text{ правильных компонент в } P) / (\# \text{ компонент в } T)$
 - Полнота = $(\# \text{ правильных компонент в } P) / (\# \text{ компонент в } P)$
 - F-мера = $2PR / (P + R)$
- Современные алгоритмы показывают точность и полноту более 90%

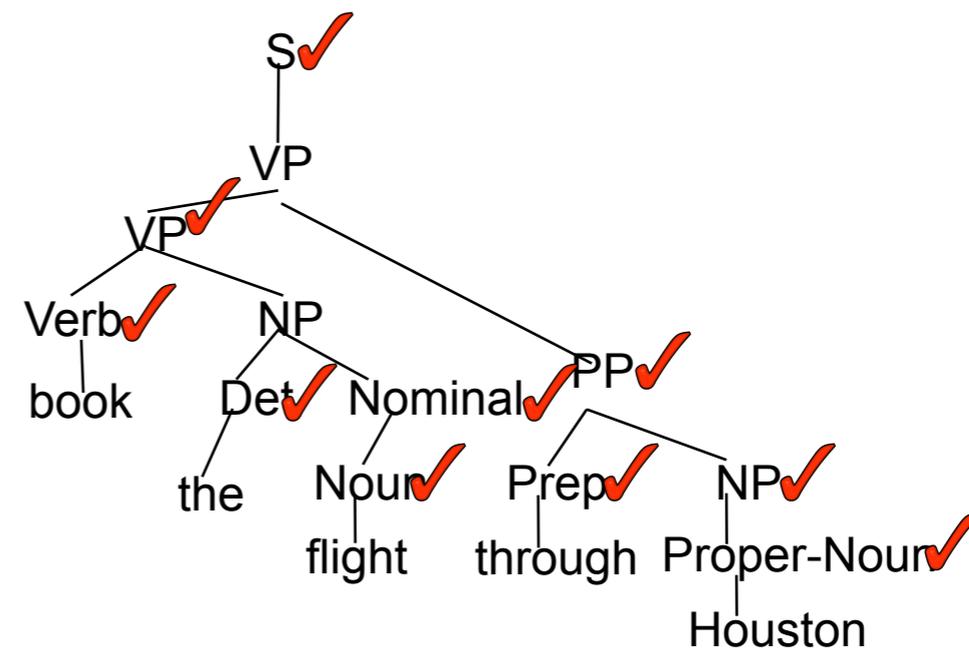
Оценка качества алгоритма

T - дерево, размеченное вручную



компонент: 12

P - вычисленное дерево



компонент: 12

правильных компонент: 10

Точность = $10/12 = 83.3\%$

Полнота = $10/12 = 83.3\%$

$F_1 = 83.3\%$

Делают ли люди синтаксический разбор?

- Психолингвистика
- Алгоритмы синтаксического разбора могут быть использованы для предсказания времени, которое потребуется человеку для прочтения каждого слова в предложении
- Чем выше вероятность слова, тем скорость чтения больше
- Для моделирования этого эффекта требуется инкрементальный алгоритм

Предложения с временной неоднозначностью

- Garden path sentence
 - Complex houses married students
 - The horse raced past the barn fell
- Инкрементальные парсеры могут найти и объяснить сложность таких предложений

Сложность языка

- Является ли естественный язык регулярным?
 - контр-пример был на прошлой лекции
- Является ли естественный язык контекстно-свободным?
 - Диалект немецкого языка в Швейцарии содержит контекстно-зависимые конструкции вида $a^n b^m c^n d^m$
- Сложность понимания людьми
 - чем проще конструкция, тем легче понимание смысла текста

Заключение

- Статистические модели, такие как СКС позволяют разрешать многозначность
- СКС можно выучить на основе банка деревьев
- Учет лексики и разделение нетерминальных символов позволяет разрешить дополнительные неоднозначности
- Точность современных алгоритмов синтаксического разбора высока, но не достигает уровня экспертного разбора

Следующая лекция

- Лексическая семантика и разрешение лексической многозначности