CSCI 4152/6509 — Natural Language Processing

2-Nov-2009

Lecture 21: Probabilistic Context-Free-Grammars

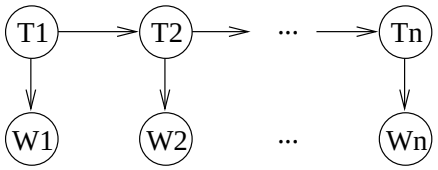
Room: FASS 2176
Time: 11:35 – 12:25

Previous Lecture

- Message passing algorithm (cont'd):
- marginalization with one variable,
- marginalization in general,
- conditioning with one variable,
- arbitrary conditional probability,
- most probable completion;
- the burglar-earthquake example

12.5 HMM as Bayesian Network

HMM Example (revisited)



Training data:

```
swat V flies N like P ants N
time N flies V like P an D arrow N
```

Generated Tables

T_1	$\mathrm{P}(T_1)$
N	0.5
V	0.5

,

T_{i-1}	T_i	$\mathrm{P}(T_i \mid T_{i-1})$
D	N	1
N	P	0.5
N	V	0.5
P	D	0.5
P	N	0.5
V	N	0.5
V	P	0.5

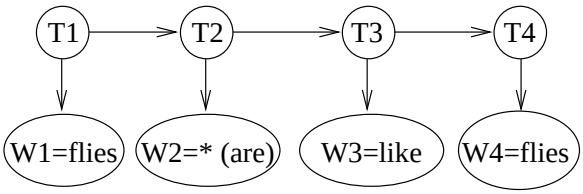
, and

T_i	W_i	$\mathrm{P}(W_i \mid T_i)$
D	an	$2/3 \approx 0.666666667$
D	*	$1/3 \approx 0.333333333$
N	ants	$2/9 \approx 0.222222222$
N	arrow	$2/9 \approx 0.222222222$
N	flies	$2/9 \approx 0.222222222$
N	time	$2/9 \approx 0.222222222$
N	*	$1/9 \approx 0.111111111$
P	like	0.8
P	*	0.2
V	flies	0.4
V	swat	0.4
V	*	0.2

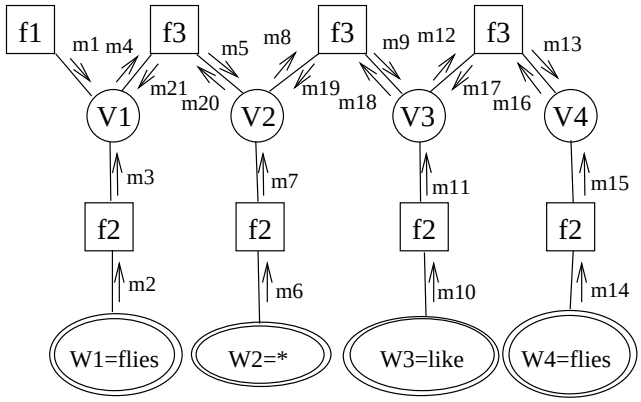
.

Tagging Example

Let us again use the example sentence: "flies are like flies"



The corresponding factor graph is:



The messages are calculated as follows:

V_1	m_1
D	0
N	0.5
P	0
V	0.5

, and

W_1	m_2
flies	1
an	0
*	0
$\vdots$	0

.

Calculation of m_3 is done as follows:

m_3			
$V_1 = D$	$W_1 =$	flies: $1 \cdot 0$	$= 0$
	$W_1 =$	an: $0 \cdot \frac{2}{3}$	$= 0$
	$W_1 =$	$\vdots$ $\vdots$	$= 0$
			max:0
$V_1 = N$	$W_1 =$	flies $: 1 \cdot \frac{2}{9}$	$= \frac{2}{9}$
	$W_1 =$	an $: 0 \cdot \frac{2}{3}$	$= 0$
			max:$2/9$
$\vdots$			

and we obtain

V_1	m_3
D	0
N	$2/9$
P	0
V	0.4

. The other messages are:

V_1	$m_4 (= m_1 \cdot m_3)$
D	$0 \cdot 0 = 0$
N	$0.5 \cdot 2/9 = 1/9$
P	$0 \cdot 0 = 0$
V	$0.5 \cdot 0.4 = 0.2$

V_2	m_5
D	0
N	0.1
P	0.1
V	$1/18$

m_5 is calculated as follows:

m_5	$m_4 \cdot f_3$	
$V_2 = D$	$V_1 = D : 0 \cdot 0$	$= 0$
	$V_1 = N : \frac{1}{9} \cdot 0$	$= 0$
	$V_1 = P : 0 \cdot 0.5$	$= 0$
	$V_1 = V : 0.2 \cdot 0$	$= 0$
		max:0

m_5	$m_4 \cdot f_3$	
$V_2 = N$	$V_1 = D : 0 \cdot 1$	$= 0$
	$V_1 = N : \frac{1}{9} \cdot 0$	$= 0$
	$V_1 = P : 0 \cdot 0.5$	$= 0$
	$V_1 = V : 0.2 \cdot 0.5$	$= 0.1$
		max:0.1

m_5	$m_4 \cdot f_3$	
$V_2 = P$	$V_1 = D : 0 \cdot 0$	$= 0$
	$V_1 = N : \frac{1}{9} \cdot 0.5$	$= 1/18$
	$V_1 = P : 0 \cdot 0$	$= 0$
	$V_1 = V : 0.2 \cdot 0.5$	$= 0.1$
		max:0.1

m_5	$m_4 \cdot f_3$	
$V_2 = V$	$V_1 = D : 0 \cdot 0$	$= 0$
	$V_1 = N : \frac{1}{9} \cdot 0.5$	$= 1/18$
	$V_1 = P : 0 \cdot 0$	$= 0$
	$V_1 = V : 0.2 \cdot 0$	$= 0$
		max:1/18

We continue calculating:

W_2	m_6
flies	0
an	0
*	1
⋮	0

,

V_2	m_7
D	$1/3$
N	$1/9$
P	0.2
V	0.2

,

V_2	$m_8 (= m_5 \cdot m_7)$
D	$0 \cdot \frac{1}{3} = 0$
N	$0.1 \cdot \frac{1}{9} = 1/90$
P	$0.1 \cdot 0.2 = 0.02$
V	$\frac{1}{18} \cdot 0.2 = 1/90$

.

To calculate m_9, we have the following intermediate calculations:

m_9	$m_8 \cdot f_3$	
$V_3 = D$	$V_2 = D : 0 \cdot 0$	$= 0$
	$V_2 = N : \frac{1}{90} \cdot 0$	$= 0$
	$V_2 = P : \frac{1}{50} \cdot 0.5$	$= 0.01$
	$V_2 = V : \frac{1}{90} \cdot 0$	$= 0$
		max:0.01

m_9	$m_8 \cdot f_3$	
$V_3 = N$	$V_2 = D : 0 \cdot 0$	$= 0$
	$V_2 = N : \frac{1}{90} \cdot 0$	$= 0$
	$V_2 = P : \frac{1}{50} \cdot 0.5$	$= 0.01$
	$V_2 = V : \frac{1}{90} \cdot 0.5$	$= 1/180$
		max:0.01

m_9	$m_8 \cdot f_3$	
$V_3 = P$	$V_2 = D : 0 \cdot 0$	$= 0$
	$V_2 = N : \frac{1}{90} \cdot 0.5$	$= 1/180$
	$V_2 = P : \frac{1}{50} \cdot 0$	$= 0$
	$V_2 = V : \frac{1}{90} \cdot 0.5$	$= 1/180$
		max:1/180

m_9	$m_8 \cdot f_3$	
$V_3 = V$	$V_2 = D : 0 \cdot 0$	$= 0$
	$V_2 = N : \frac{1}{90} \cdot 0.5$	$= 1/180$
	$V_2 = P : \frac{1}{50} \cdot 0$	$= 0$
	$V_2 = V : \frac{1}{90} \cdot 0$	$= 0$
		max:1/180

and we obtain:

V_3	m_9
D	0.01
N	0.01
P	1/180
V	1/180

. Then,

W_3	m_{10}
like	1
⋮	0

,

V_3	m_{11}
D	0
N	0
P	0.8
V	0

,

V_3	$m_{12} (= m_9 \cdot m_{11})$
D	$0.01 \cdot 0 = 0$
N	$0.01 \cdot 0 = 0$
P	$\frac{1}{180} \cdot 0.8 = 1/225$
V	$\frac{1}{180} \cdot 0 = 0$

.

To calculate m_{13}, we have the following intermediate calculations:

m_{13}	$m_{12} \cdot f_3$	
$V_3 = D$	$V_2 = D : 0 \cdot 0$	$= 0$
	$V_2 = N : 0 \cdot 0$	$= 0$
	$V_2 = P : \frac{1}{225} \cdot 0.5$	$= 1/450$
	$V_2 = V : 0 \cdot 0$	$= 0$
		max:1/450

m_{13}	$m_{12} \cdot f_3$	
$V_3 = N$	$V_2 = D : 0 \cdot 1$	$= 0$
	$V_2 = N : 0 \cdot 0$	$= 0$
	$V_2 = P : \frac{1}{225} \cdot 0.5$	$= 1/450$
	$V_2 = V : 0 \cdot 0.5$	$= 0$
		max:1/450

m_{13}	$m_{12} \cdot f_3$	
$V_3 = P$	$V_2 = D : 0 \cdot 0$	$= 0$
	$V_2 = N : 0 \cdot 0.5$	$= 0$
	$V_2 = P : \frac{1}{225} \cdot 0$	$= 0$
	$V_2 = V : 0 \cdot 0.5$	$= 0$
		max:0

m_{13}	$m_{12} \cdot f_3$	
$V_3 = V$	$V_2 = D : 0 \cdot 0$	$= 0$
	$V_2 = N : 0 \cdot 0.5$	$= 0$
	$V_2 = P : \frac{1}{225} \cdot 0$	$= 0$
	$V_2 = V : 0 \cdot 0$	$= 0$
		max:0

and we obtain:

V_4	m_{13}
D	$1/450$
N	$1/450$
P	0
V	0

. Then,

W_4	m_{14}
flies	1
$\vdots$	0

, and

V_4	m_{15}
D	0
N	$2/9$
P	0
V	0.4

.

To maximize the product of probabilities of V_4 we calculate:

V_4	$m_{13} \cdot m_{15}$	
D	$\frac{1}{450} \cdot$	$= 0$
N	$\frac{1}{450} \cdot \frac{2}{9}$	$= 1/2025$
P	$0 \cdot 0$	$= 0$
V	$0 \cdot 0.4$	$= 0$

and we obtain $\boxed{V_4^* = N}$, which we use in further messages, as a "hard-wired" value. We calculate

V_4	m_{16}
D	0
N	$2/9$
P	0
V	0

, and for m_{17} use only $V_4 = N$ in $m_{16} \cdot f_3$:

$m_{16} \cdot f_3$	
$\frac{2}{9} \cdot 1$	$= 2/9$
$\frac{2}{9} \cdot 0$	$= 0$
$\frac{2}{9} \cdot 0.5$	$= 1/9$
$\frac{2}{9} \cdot 0.5$	$= 1/9$

, and we obtain:

V_3	m_{17}
D	$2/9$
N	0
P	$1/9$
V	$1/9$

.

To find optimal V_3 we calculate:

V_3	$m_9 \cdot m_{11} \cdot m_{17}$	
D	$0.01 \cdot 0 \cdot \frac{2}{9}$	$= 0$
N	$0.01 \cdot 0 \cdot 0$	$= 0$
P	$\frac{1}{180} \cdot 0.8 \cdot \frac{1}{9}$	$= 1/2025$
V	$\frac{1}{180} \cdot 0 \cdot \frac{1}{9}$	$= 0$

and we obtain $\boxed{V_3^* = P}$.

Then,

V_3	$m_{18} = m_{17} \cdot m_{11}$
D	0
N	0
P	$\frac{1}{9} \cdot 0.8 = 4/45$
V	0

,

V_2	$m_{19} = m_{18} \cdot f_3$ for $V_3 = P$
D	$\frac{4}{45} \cdot 0 = 0$
N	$\frac{4}{45} \cdot \frac{1}{2} = 2/45$
P	$\frac{4}{45} \cdot 0 = 0$
V	$\frac{4}{45} \cdot \frac{1}{2} = 2/45$

.

To find optimal V_2 we calculate:

V_2	$m_{19} \cdot m_5 \cdot m_7$	
D	$0 \cdot 0 \cdot \frac{1}{3}$	$= 0$
N	$\frac{2}{45} \cdot 0.1 \cdot \frac{1}{9}$	$= 1/2025$
P	$0 \cdot 0.1 \cdot 0.2$	$= 0$
V	$\frac{2}{45} \cdot \frac{1}{18} \cdot 0.2$	$= 1/2025$

and we can choose either N or V. Let us choose $\boxed{V_2^* = V}$.

V_2	$m_{20} = m_7 \cdot m_{19}$
D	0
N	0
P	0
V	$0.2 \cdot \frac{2}{45} = 2/225$

,

V_1	$m_{21} = m_{20} \cdot f_3$ for $V_2 = V$
D	$\frac{2}{225} \cdot 0 = 0$
N	$\frac{2}{225} \cdot \frac{1}{2} = 1/225$
P	$\frac{2}{225} \cdot 0 = 0$
V	$\frac{2}{225} \cdot 0 = 0$

.

To find optimal V_1 we calculate:

V_1	$m_1 \cdot m_3 \cdot m_{21}$	
D	$0 \cdot 0 \cdot 0$	$= 0$
N	$0.5 \cdot \frac{2}{9} \cdot \frac{1}{225}$	$= 1/2025$
P	$0.5 \cdot 0 \cdot 0$	$= 0$
V	$0 \cdot 0.4 \cdot 0$	$= 0$

and we obtain $\boxed{V_1^* = N}$.

13 Probabilistic Context-Free Grammar (PCFG)

Reading: Chapters 13 and 14

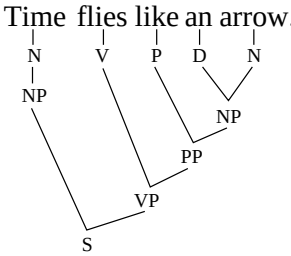
Probabilistic Context-Free Grammar (PCFG) is also known as **Stochastic Context-Free Grammar (SCFG)**. Both, n-gram model and HMM are linear models, which may not be most suitable to model the structured nature of natural language syntax. While Bayesian Networks could be one way of capturing structured nature of language in a probabilistic way, PCFGs represent another way, which is directly derived from the Context-Free Grammar formalism.

For example, in language modelling applied to the sentence:

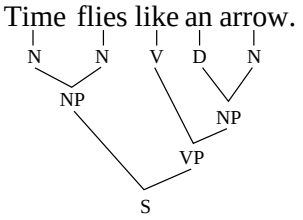
The velocity of the seismic waves rises to...

a linear model will likely assign a higher probability to the word "rise" after the plural "waves" than to the word "rises," which actually correctly appears in the sentence and agrees with the head "velocity" of the noun phrase.

As previously described, context-free grammars represent a structural model for describing syntax. For example, the syntax of the sentence "Time flies like an arrow." could be represented as the following context-free parse tree:



There are known efficient parsing algorithms for context-free grammars in the theory of of formal languages, and applications such as design of compilers and interpreters for programming languages. Two examples of such parsing approaches are recursive descent parsing and shift-reduce LR parsing. A large obstacle in applying these parsers to the problem of NL parsing is in the requirement that the language is unambiguous. Natural languages are inherently ambiguous and a parser for natural language must handle ambiguous grammars and ambiguous input. For example, if we assume a different meaning of the above sentence, we obtain a different parse tree, like the following one:



The above two trees induce the following CFG:

S	→	NP VP	VP	→	V NP	N	→	time	V	→	like
NP	→	N	VP	→	V PP	N	→	arrow	V	→	flies
NP	→	N N	PP	→	P NP	N	→	flies	P	→	like
NP	→	D N				D	→	an			

To have a complete CFG specification, we need to add that the set of terminals is {'time', 'arrow', 'flies', 'an', 'like'}, the set of non-terminals is { S, NP, VP, D, N, PP, P, V}, and the start symbol is S.

If we parse the same sentence using this grammar, then we will obtain at least two different parse trees. To make parsing more usable, we need a way of assigning a score or probability to each tree, so we can always choose the "best" parse tree in a certain sense.

13.1 PCFG as a Probabilistic Model

To transform a CFG into a probabilistic model we model derivations as stochastic process in a generative way. For example, the left-most derivation corresponding to the first parse tree described above is:

$$\begin{aligned} \text{S} &\Rightarrow \text{NP VP} \Rightarrow \text{N VP} \Rightarrow \text{time VP} \Rightarrow \text{time V PP} \Rightarrow \text{time flies PP} \Rightarrow \text{time flies P NP} \\ &\Rightarrow \text{time flies like NP} \Rightarrow \text{time flies like D N} \Rightarrow \text{time flies like an N} \Rightarrow \text{time flies like an arrow} \end{aligned}$$

At each step of the derivation, given a non-terminal that needs to be re-written, we usually have several options, corresponding to several rules that have this non-terminal on the left-hand side.
Hence, we calculate the probability of the tree by multiplying probabilities of all rules occurring in the tree:

$$\begin{aligned} \mathrm{P}(\text{first tree}) &= \mathrm{P}(\text{N} \rightarrow \text{time})\mathrm{P}(\text{V} \rightarrow \text{flies})\mathrm{P}(\text{P} \rightarrow \text{like})\mathrm{P}(\text{D} \rightarrow \text{an}) \\ &\quad \mathrm{P}(\text{N} \rightarrow \text{arrow})\mathrm{P}(\text{NP} \rightarrow \text{N})\mathrm{P}(\text{NP} \rightarrow \text{D N}) \ldots \mathrm{P}(\text{S} \rightarrow \text{NP VP}) \end{aligned}$$

If we assign the following probabilities to the rules:

S	→	NP VP	/1	VP	→	V NP	/.5	N	→	time	/.5
NP	→	N	/.4	VP	→	V PP	/.5	N	→	arrow	/.3
NP	→	N N	/.2	PP	→	P NP	/1	N	→	flies	/.2
NP	→	D N	/.4					D	→	an	/1
V	→	like	/.3								
V	→	flies	/.7								
P	→	like	/1								

then the probability of the first tree is 0.0084, and the probability of the second tree is 0.00036. We can conclude that the first tree is more likely, which should correspond to our intuition.
The probability assigned to a rule $N \rightarrow \alpha$ is the probability $\mathrm{P}(N \rightarrow \alpha | N)$, so if $N \rightarrow \alpha_1$, $N \rightarrow \alpha_2$, …, $N \rightarrow \alpha_n$ are all rules with the nonterminal N on its left hand side, then

$$\sum_{i=1}^{n} \mathrm{P}(N \rightarrow \alpha_i) = 1$$

These probabilities are easily learned from a set of parse trees, usually called parse treebank, by counting the number of occurrences of distinct rules.
This model is a language model, since the sum of probabilities of all sentences in the language is 1. Actually, in order to be a language model, we also require that the grammar is *proper,* i.e., that all infinite trees have probability 0, which is not always the case. We will not go into further details regarding this question here, except noting that it has been proved that any PCFG with probabilities induced from a treebank is proper.