SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 98
Accredited by NAAC-UGC with ‘A’ Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – 19AD601 – Natural Language Processing

III Year / VI Semester

Unit 3 – SYNTACTIC ANALYSIS

Topic 7- Shallow parsing – Probabilistic CFG

Shallow parsing – Probabilistic CFG

- The shallow parser gives the analysis of a sentence in terms of morphological analysis, POS tagging, Chunking, etc. Apart from the final output, intermediate output of individual modules is also available.

- Chunking (aka. Shallow parsing) is to analyzing a sentence to identify the constituents (noun groups, verbs, verb groups, etc.). However, it does not specify their internal structure, nor their role in the main sentence.

- Shallow parsing, also known as light parsing or chunking , is a popular natural language processing technique of analyzing the structure of a sentence to break it down into its smallest constituents (which are tokens such as words) and group them together into higher-level phrases.

- This includes POS tags as well as phrases from a sentence.

Shallow parsing – Probabilistic CFG

Probabilistic CFG

- PCFG is a simple extension of a CFG in which every production rule is associated with a probability.
- A PCFG consists of:
- 1. A context-free grammar G = (N, Σ, S, R).
- 2. 2. A parameter
- q(α → β)
- for each rule α → β ∈ R. The parameter q(α → β) can be interpreted as the conditional probabilty of choosing rule α → β in a left-most derivation, given that the non-terminal being expanded is α. For any X ∈ N , we have the constraint.

$$\sum_{\alpha\rightarrow\beta\in R:\alpha=X} q(\alpha\rightarrow\beta) = 1$$

Shallow parsing – Probabilistic CFG

- Given a parse-tree t ∈ TG containing rules α1 → β1, α2 → β2, . . . , αn → βn, the probability of t under the PCFG is

$$p(t) = \prod_{i=1}^{n} q(\alpha_i \rightarrow \beta_i)$$

The only addition to the original context- free grammar is a parameter q(α → β) for each rule α → β ∈ R. Each of these parameters is constrained to be non-negative, and in addition we have the constraint that for any non-terminal X ∈ N.

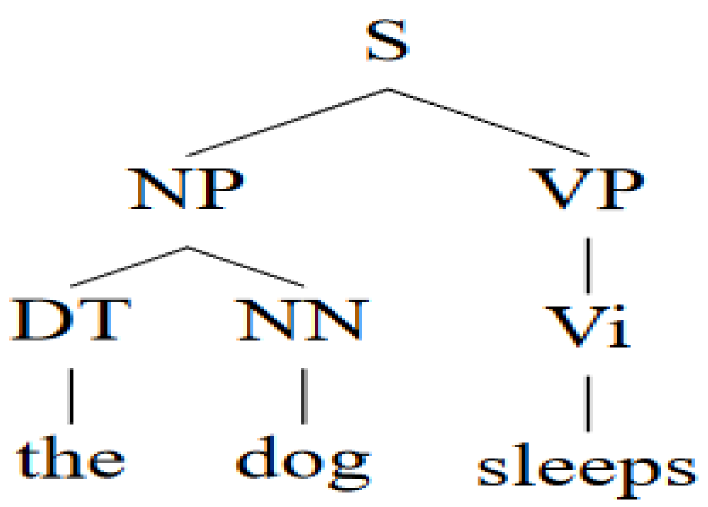
$$\sum_{\alpha \rightarrow \beta \in R : \alpha = X} q(\alpha \rightarrow \beta) = 1$$

Shallow parsing – Probabilistic CFG

- This simply states that for any non-terminal X, the parameter values for all rules with that non-terminal on the left-hand-side of the rule must sum to one,

$$\begin{aligned}\sum_{\alpha\rightarrow\beta\in R:\alpha=\mathtt{VP}} q(\alpha\rightarrow\beta) &= q(\mathtt{VP}\rightarrow\mathtt{Vi}) + q(\mathtt{VP}\rightarrow\mathtt{Vt\ \ NP}) + q(\mathtt{VP}\rightarrow\mathtt{VP\ \ PP}) \\ &= 0.3+0.5+0.2 \\ &= 1.0\end{aligned}$$

To calculate the probability of any parse tree t, we simply multiply together the q values for the context-free rules that it contains. For example, if our parse tree t is



Shallow parsing – Probabilistic CFG

- then we have p(t) = q(S → NP VP) × q(NP → DT NN) × q(DT → the) × q(NN → dog) ×
- q(VP → Vi) × q(Vi → sleeps)
- Intuitively, PCFGs make the assumption that parse trees are generated stochastically, according to the following process:9

N = {S, NP, VP, PP, DT, Vi, Vt, NN, IN}
S = S
Σ = {sleeps, saw, man, woman, dog, telescope, the, with, in}

$R, q =$

S	→	NP	VP	1.0
VP	→	Vi		0.3
VP	→	Vt	NP	0.5
VP	→	VP	PP	0.2
NP	→	DT	NN	0.8
NP	→	NP	PP	0.2
PP	→	IN	NP	1.0

Vi	→	sleeps	1.0
Vt	→	saw	1.0
NN	→	man	0.1
NN	→	woman	0.1
NN	→	telescope	0.3
NN	→	dog	0.5
DT	→	the	1.0
IN	→	with	0.6
IN	→	in	0.4

Shallow parsing – Probabilistic CFG

- then we have p(t) = q(S → NP VP) × q(NP → DT NN) × q(DT → the) × q(NN → dog) ×
- q(VP → Vi) × q(Vi → sleeps)
- Intuitively, PCFGs make the assumption that parse trees are generated stochastically, according to the following process:9

$N = \{\text{S, NP, VP, PP, DT, Vi, Vt, NN, IN}\}$
$S = \text{S}$
$\Sigma = \{\text{sleeps, saw, man, woman, dog, telescope, the, with, in}\}$

$R, q =$

S	→	NP	VP	1.0
VP	→	Vi		0.3
VP	→	Vt	NP	0.5
VP	→	VP	PP	0.2
NP	→	DT	NN	0.8
NP	→	NP	PP	0.2
PP	→	IN	NP	1.0

Vi	→	sleeps	1.0
Vt	→	saw	1.0
NN	→	man	0.1
NN	→	woman	0.1
NN	→	telescope	0.3
NN	→	dog	0.5
DT	→	the	1.0
IN	→	with	0.6
IN	→	in	0.4

THANK YOU