Probabilistic Parsing کمش

④ از رابطه‌ی روبرو استفاده می‌کنیم ←

$$\forall X \in N, \sum_{X \to \gamma \in R} P(X \to \gamma) = 1$$

برای VP : $X1 + 0.5 = 1 \rightarrow X1 = 0.5$

برای NP : $0.2 + 0.6 + X2 = 1 \rightarrow X2 = 0.2$

برای PP : $X3 = 1$

برای N : $0.4 + 0.3 + X4 + 0.1 = 1 \rightarrow X4 = 0.2$

برای V : $0.2 + 0.6 + X5 = 1 \rightarrow X5 = 0.2$

⑤ P(t) = the probability of a tree t is the product of the probabilities of the rules used to generate it.

P(s) = the probability of the string s is the sum of the probabilities of the trees which have that string as their yield.

$P(s) = \sum_j P(s,t)$ where t is a parse of $s = \sum_j P(t)$

VP→ VP PP (0.3) : احتمالاً "rule" صورت سوال غلط درنوشتن است ←

درخت سمت چپ :

$P(t_1) = 1.0 \times 0.3 \times 0.4 \times 0.7 \times 1 \times 0.3 \times 0.5 \times 0.3$

S NP N VP V NP N VP

$\times 1.0 \times 1 \times 0.2 \times 0.3 \times 0.4 \times 1 \times 1 \times 0.3 \times 0.1 = 0.0000027216$

PP P NP NP N PP P NP N

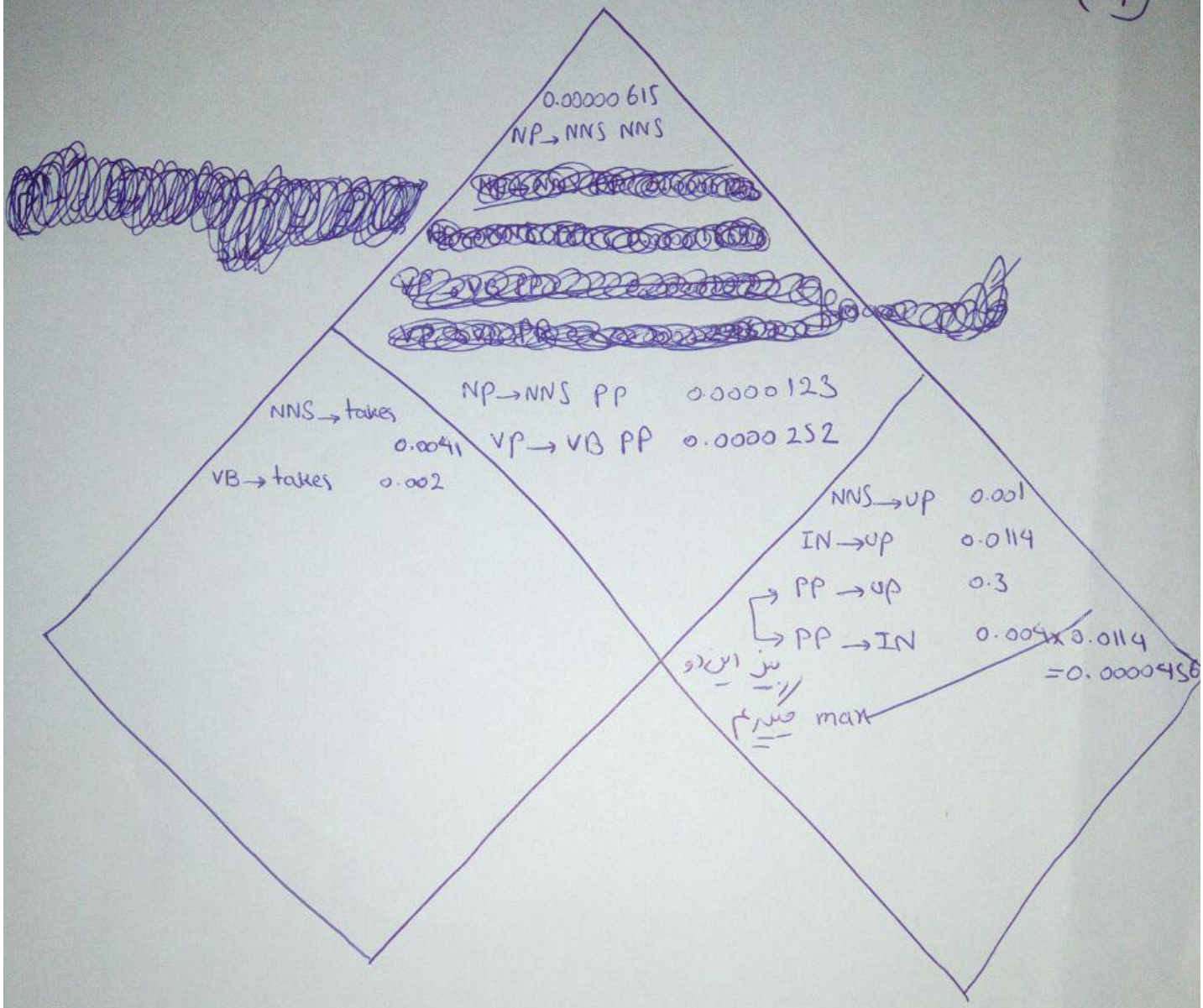
S NP N VP V NP NP N PP P NP NP PP P NP

$P(t_2) = 1 \times 0.3 \times 0.4 \times 0.7 \times 1 \times 0.2 \times 0.3 \times 0.5 \times 1 \times 1 \times 0.2 \times 0.3 \times 1 \times 1 \times 0.3$

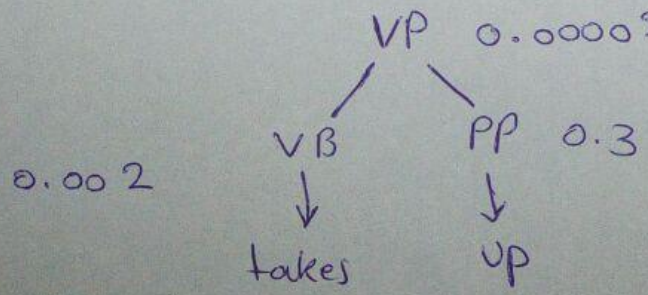
N N

$0.4 \times 0.1 = 0.0000018144$

$P(s) = P(t_1) + P(t_2) = 0.00000 4536$



← از بالا به پایین با توجه به بیشترین احتمال rule های ممکن عبارت داده شده را پارس می‌کنیم:



فصل ١١

(١)

5.2.1 problems with PCFG:

- The problems in structural dependency

A CFG assumes that the expansion of any one non-terminal is independent of the expansion of any other non-terminal. This independence assumption is carried over in the PCFG: each PCFG rule is assumed to be independent of each other rule, and thus the rule probabilities are multiplied together. But, In English, the choice of how a node expands is dependent on the location of the node in the parse tree. For example, there is a strong tendency for the syntactic subject of a sentence to be a pronoun. This tendency is caused by the use of subject position to realize the topic or old information. Pronouns are a way to talk about old information. While the non-pronominal lexical noun-phrase are often used to introduce new referents. According to the investigation of Francis (1999), the 31,021 subjects of declarative sentences in Switchboard corpus, 91% are pronouns and only 9% are lexical. By contrast, out of 7,498 direct object, only 34% are pronoun, and 66% are lexical.

Subject: **She** is able to take her baby to work with her.

: **My wife** worked until we had a family.

Object: Some laws absolutely prohibit **it**.

All the people signed **applications**.

These dependencies could be captured if the probability of expanding an NP as a pronoun (via the rule NP → Pronoun) versus a lexical NP (via the rule NP → Det Noun) were dependent on whether the NP was a subject or an object. However, this is just the kind of probabilistic dependency that a PCFG does not allow.

■ The problems in lexical dependency

PCFG can only be represented via the probability of pre-terminal nodes to be expanded lexically. But there are a number of other kinds of lexical and other dependencies that is important in modeling syntactic probabilities.

--PP-attachment: The lexical information plays an important role in selecting the correct parsing of an ambiguous prepositional phrase attachment.

For example, in the sentence "Washington sent more than 10,000 soldiers into Afghanistan", PP "into Afghanistan" can be attached either to NP (more than 10,000 soldiers), or to attached to the verb (sent).

In PCFG, the attachment choice comes down to the choice between two rules:

NP → NP PP (NP-attachment)

And VP → VP PP (VP-attachment)

The probability of these two rules depends on the training corpus.

Corpus	NP-attachment	VP-attachment
AP Newswire (13 million words)	67%	33%
Wall Street Journal & IBM manuals	52%	48%

Whether the preference is 67% or 52%, in PCFG, this preference is purely structural and must b the same to all verb.

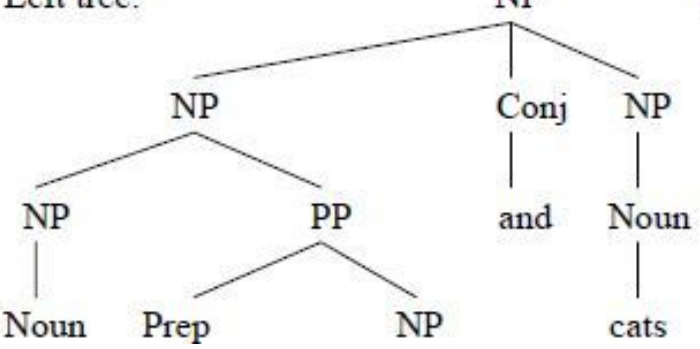
However, the correct attachment is to verb. The verb "send" subcategorizes for a destinatior which can be expressed with the preposition "into". It is a lexical dependency. The PCFG can nc deal with the lexical dependency.

--Coordination ambiguities:

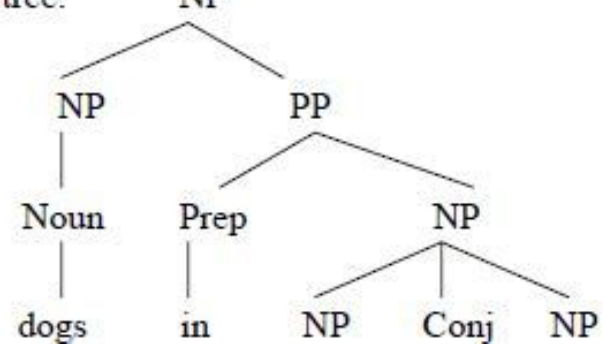
The coordination ambiguities are the key to choosing the proper parse.

In the phrase "dogs in houses and cats" is ambiguous:

Left tree:

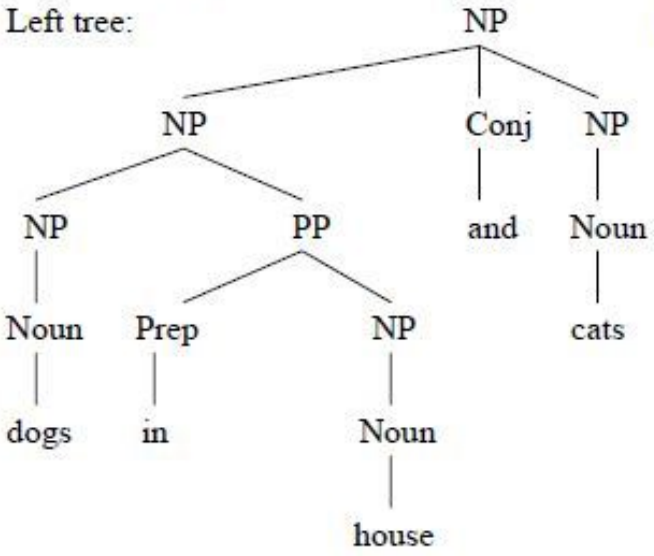


Right tree:

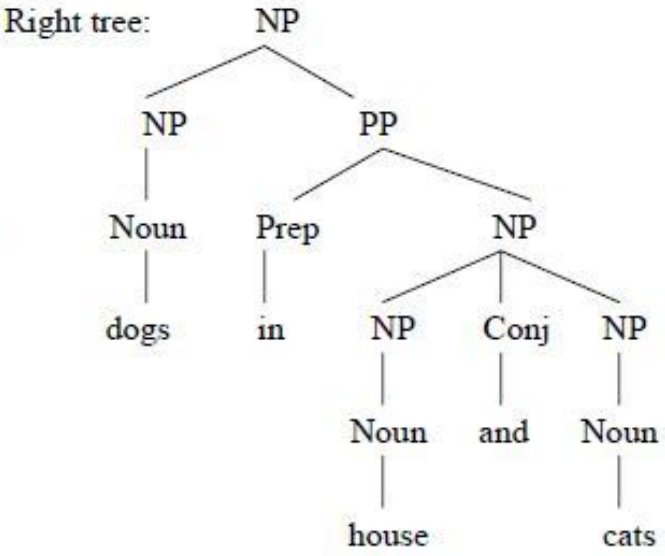


In the phrase "dogs in houses and cats" is ambiguous:

Left tree:



Right tree:



Although the left tree is intuitively the correct one. But the PCFG will assign them identically probabilities because both structure use the exact same rule:

NP → NP Conj NP

NP → NP PP

NP → Noun

PP → Prep NP

Noun → dogs | house | cats

Prep → in

Conj → and

In this case, PCFG will assign two trees the same probability.

PCFG has a number of inadequacies as a probabilistic model of syntax, we shall augment PCFG to deal with these problems.

در این مقاله در روش lexicalization با اضافه کردن lexical information به هر rule

حل می شود انتخاب parser بهتر و دقیق تری انجام می شود.