PCFG PARSING II

CS662: Natural Language Processing

2015-02-24

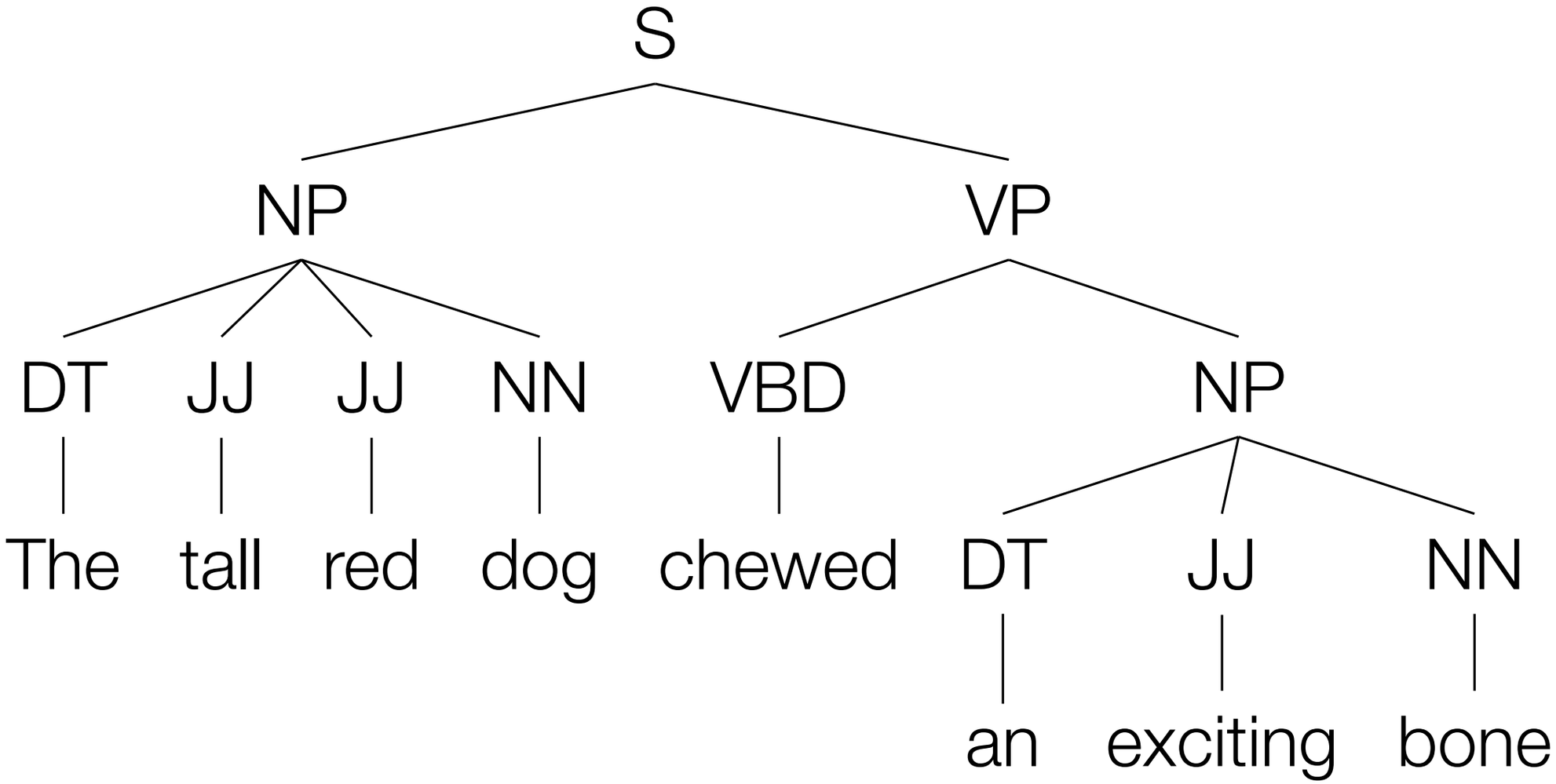
Penn Treebank non-terminal labels (LHSes) are relatively simple and do not contain much information about the context they occur in.

Penn Treebank constituents (RHSes) are relatively flat (i.e., long) and thus rather sparse.

Thus PCFG rules of the form

$$p(\text{LHS} \rightarrow \text{RHS}) = p(\text{RHS} \mid \text{LHS})$$

are too uniform in LHS and too sparse in RHS.


S
NP
VP
DT
JJ
JJ
NN
VBD
NP
The
tall
red
dog
chewed
DT
JJ
NN
an
exciting
bone

Klein & Manning (2003) identify two major factors that determine how PCFG rules can be conditioned:

The degree of vertical context of the LHS, *v*

The degree of horizontal context in the RHS, *h*

In PTB, $v = 1$ and $h = \infty$. (This is historical accident.)

It is natural to add LHS vertical context by increasing *v*, and reduce RHS sparsity by decreasing *h*

TAG SPLITTING

Penn Treebank tags are not sufficiently fine-grained to capture important $p(w \mid t)$ distinctions, so we can decorate preterminals with their parents.

E.g., all the following are tagged `IN`:

Subordinating conjunctions (*while*, *as*, *if*), under `IN^S`

Complementizers (*that*, *for*) under `IN^SBAR`

Prepositions (*of*, *in*, *from*) under `IN^PP`

"…the most common adverbs [**RB**—KG] directly under **ADVP** are *also* (1599) and *now* (544). Under **VP**, they are *n't* (3779) and *not* (922). Under **NP**, *only* (215) and *just* (132), and so on."
(Klein & Manning 2003)

PARENT ANNOTATION

For $v = 2$, it is not merely

$$p(\texttt{DT JJ JJ NN} \mid \texttt{NP})$$

but now

$$p(\texttt{DT JJ JJ NN} \mid \texttt{NP\^S}) \,.$$

So, e.g., **NP**s with **S** parents (subjects) will be marked **NP^S** whereas **NP**s with **VP** parents (objects) will be **NP^VP**.

“The category symbols are too coarse to adequately render the expansions independent of the contexts. For example, subject **NP** expansions are very different from object **NP** expansions: a subject **NP** is 8.7 times more likely than an object **NP** to expand as just a pronoun. Having separate symbols for subject and object **NP**s allows this variation to be captured and used to improve parse scoring.” (Klein & Manning 2003)

HEAD ANNOTATION

First, condition the LHS (e.g., S) on the head child category *H* (e.g., VP) in the RHS; e.g., S(VP).

Then, binarize the non-head categories of the RHS to the left (e.g., ADVP) and right (VBZ, NP, PP) and generate them working from the inside out.

LEXICALIZATION

Condition LHSes (e.g., **NP**) based on the head word H_w (e.g., **dog**) of the RHS; e.g., **NP(dog)**, rather than just the category of the head.

COLLINS PARSER MODEL

Essentially: $v = 2$, $h = 2$, with lexicalization.

To expand a non-terminal category P:

First, pick the category of the RHS head H according to P, the headword terminal H_w, and headword tag H_t:

$$p(H \mid P, H_w, H_t)$$

...

COLLINS PARSER MODEL

Then, to generate each child, pick the child's category C and headword tag CH_t according to P, H, H_w, H_t, and the distance between the child and the RHS head Δ:

$$p_c(C, CH_t \mid P, H, H_w, H_t, \Delta)$$

and then pick the child's new headword according to:

$$p_{cw}(CH_w \mid P, H, H_w, H_t, \Delta, C, CH_t) \ .$$

BILEXICAL DEPENDENCIES

Collins conditions the selection of constituents heads CH_w of the RHS on the headword H_w of the constituent.

But, Gildea (2001) finds that bilexical dependencies add little, and don't generalize well to out-of-domain data.

Some Penn Treebank non-terminal contain information about the presence of empty categories (e..g, null subjects in relative clauses like *the dog* [*Eleanor likes*]), which is usually discarded. These non-terminals have drastically different selectional properties, so it is helpful to, e.g., relabel them `GAPPED-S`.