FNLP Tutorial 5

1 Attachment Ambiguity in Trees

A common source of attachment ambiguity in English comes from prepositional phrases. The relevant grammar rules include:

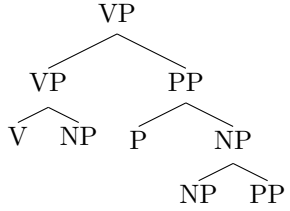
$$\text{VP} \rightarrow \text{V NP}$$
$$\text{VP} \rightarrow \text{VP PP}$$
$$\text{NP} \rightarrow \text{NP PP}$$
$$\text{PP} \rightarrow \text{P NP}$$

Here are five verb phrases:

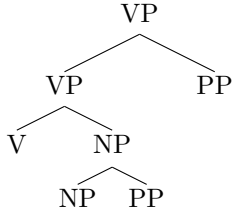
1. read the paper about the Turing test on my laptop
2. read the paper about the success of Transformer
3. read the paper by Alan Turing about the Turing test
4. joined the meeting in the office with the blue door
5. joined the online meeting from my office via Zoom

Pictured below are five partial trees. Match the phrases to the trees which best capture their meanings. Is there a source of ambiguity with the first sentence?

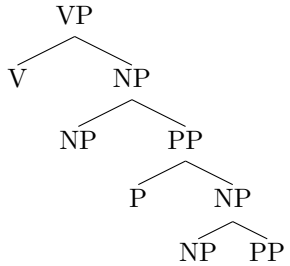
(A)



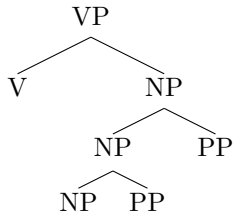
(B)



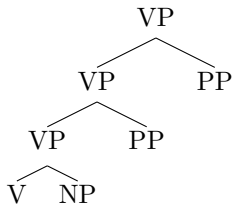
(C)



(D)



(E)



Solution 1B; 2C; 3D; 4A; 5E The first sentence could also be C, if we understood that the paper was about a Turing test done on my laptop. However, this is semantically implausible -we know Turing tests are not done on laptops-.

2 Writing a context-free grammar

Consider this small corpus of English noun phrases.

1. this book
2. the great green dragon
3. these green leaves
4. the leaves in the hallway

- Write a context-free grammar that generates at least the above noun phrases. Use the following part-of-speech tags as pre-terminal symbols: determiner (DT), adjective (A), preposition (P) and noun (N).
- Show a derivation tree of a grammatical noun phrase that is not included in the corpus. Does your grammar overgenerate? That is, does it generate sequences that are not grammatical English noun phrases? If so, refine your grammar to fix the issue that you mentioned.

Solution We will start with the following set of POS tags:

DT determiner, e.g. this and these

A adjective, e.g. great and green

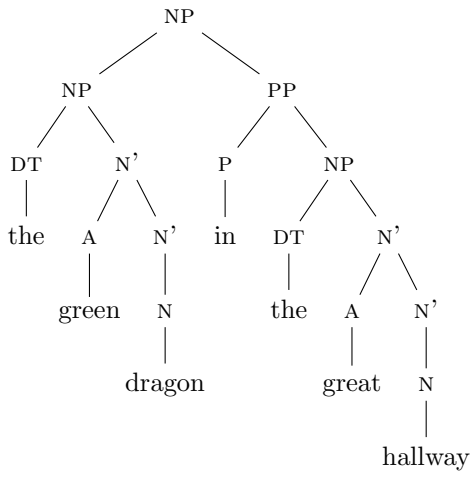
P preposition, e.g. in

N noun, e.g. book

Our first attempt might look something like this (NP is the start symbol). We write $X \to Y|Z$ as a shorthand for the two productions $X \to Y$ and $X \to Z$.

$$
\begin{aligned}
\text{NP} &\to \text{DT} \quad \text{N'} \\
\text{NP} &\to \text{NP} \quad \text{PP} \\
\text{N'} &\to \text{A} \quad \text{N'} \\
\text{N'} &\to \text{N} \\
\text{PP} &\to \text{P} \quad \text{NP} \\
\text{DT} &\to \text{the}|\text{this}|\text{these} \\
\text{P} &\to \text{in} \\
\text{A} &\to \text{great}|\text{green} \\
\text{N} &\to \text{book}|\text{dragon}|\text{leaves}|\text{hallway}
\end{aligned}
$$

We can derive the noun phrase "the green dragon in the great hallway" from this grammar:



Unfortunately, this grammar also generates ungrammatical noun phrases like:

- *this leaves
- *these great book

To prevent this from happening, our grammar must encode number information (if something is singular or plural). In particular, we need to encode this information for determiners and nouns and make sure that only singular determiners and singular nouns can combine (and analogously for plural).

We first refine our set of POS tags slightly to encode number information, i.e. we introduce DT_{SG}, DT_{PL} and N_{SG} and N_{PL}. We also modify production rules to track this information:

$$
\begin{aligned}
\text{NP} &\rightarrow \text{DT}_{\text{SG}} \quad \text{N'}_{\text{SG}} \\
\text{NP} &\rightarrow \text{DT}_{\text{PL}} \quad \text{N'}_{\text{PL}} \\
\text{NP} &\rightarrow \text{NP} \quad \text{PP} \\
\text{N'}_{\text{SG}} &\rightarrow \text{A} \quad \text{N'}_{\text{SG}} \\
\text{N'}_{\text{PL}} &\rightarrow \text{A} \quad \text{N'}_{\text{PL}} \\
\text{N'}_{\text{SG}} &\rightarrow \text{N}_{\text{SG}} \\
\text{N'}_{\text{PL}} &\rightarrow \text{N}_{\text{PL}} \\
\text{PP} &\rightarrow \text{P} \quad \text{NP} \\
\text{DT}_{\text{SG}} &\rightarrow \text{the}|\text{this} \\
\text{DT}_{\text{PL}} &\rightarrow \text{these} \\
\text{P} &\rightarrow \text{in} \\
\text{A} &\rightarrow \text{great}|\text{green} \\
\text{N}_{\text{SG}} &\rightarrow \text{book}|\text{dragon}|\text{hallway} \\
\text{N}_{\text{PL}} &\rightarrow \text{leaves}
\end{aligned}
$$

This grammar still generates noun phrases that are disfavoured if not ungrammatical, like:

- ??the green great dragon

3 The CKY Algorithm

1. Determine the potential parses of the sentence "The women fish with bait" according to the PCFG below, using the CKY algorithm. Number the symbols you put in the chart in the

order they would be computed. For each chart item, assume that candidate grammar rules are searched in the order they are listed below. Use these numbers to construct backpointers in the chart. For example, when constructing the 6th symbol NP out of a determiner with symbol number 2 and a noun with symbol number 3, you could indicate that by putting '6: NP (2, 3)'. (This is not a widely used custom, but a useful one when doing CKY on paper for the context of this tutorial.)

2. What are the probabilities of the parse trees?

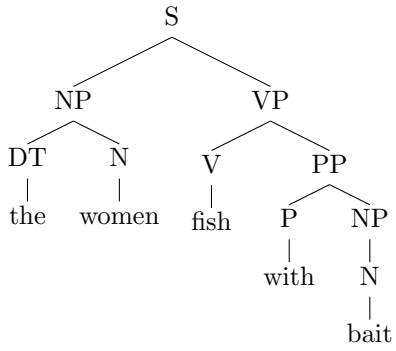
$$
\begin{aligned}
\text{S} &\rightarrow \text{NP} \quad \text{VP} \ (1) \\
\text{NP} &\rightarrow \text{DT} \quad \text{N} \ (0.5) \\
\text{NP} &\rightarrow \text{N} \quad \text{N} \ (0.1) \\
\text{NP} &\rightarrow \text{N} \ (0.4) \\
\text{VP} &\rightarrow \text{V} \quad \text{PP} \ (1) \\
\text{PP} &\rightarrow \text{P} \quad \text{NP} \ (1) \\
\text{DT} &\rightarrow \text{the} \ (1) \\
\text{P} &\rightarrow \text{with} \ (1) \\
\text{V} &\rightarrow \text{fish} \ (1) \\
\text{N} &\rightarrow \text{women} \ (0.5) \mid \text{fish} \ (0.3) \mid \text{bait} \ (0.2)
\end{aligned}
$$

Solution

1. Here is the CKY chart:

the	women	fish	with	bait
1: DT p=1	10: NP(1,2) p=0.25			15: S(10,13) p=0.02
	2: N p=0.5, 3: NP(2) p=0.2	11: NP(2,5) p=0.015		14: S(3,13) p=0.016
		4: V p=1, 5: N p=0.3, 6: NP(5) p=0.12		13: VP(4,12) p=0.08
			7: P p=1	12: PP(7,9) p=0.08
				8: N p=0.2, 9: NP(8) p=0.08

The resulting parse looks like this:



2. There is only one valid parse. The parse tree has the following probability: $1 \times 0.5 \times 1 \times 1 \times 0.2 \times 0.5 \times 0.4 \times 1 \times 1 \times 1 = 0.02$.