Accelerated Natural Language Processing 2018

Lecture 11: Grammar and grammar formalisms

Henry S. Thompson
Drawing on slides by Philip Koehn and Jurafsky and Martin 2009
9 October 2018

1. Parsing

If trees are useful, how do we get them?

Parsing is the process of taking a string and a grammar and returning one or more parse trees for that string

Analogous to running a finite-state transducer over a tape

- However, since CFGs are more powerful
  - That is, there are languages we can capture with CFGs that we can’t capture with finite-state methods
- The parsing process is likewise more complicated
  - As we'll see in a few days

2. Exploring syntax

By grammar, or syntax, we have in mind the kind of implicit knowledge of your native language that you had mastered by the time you were 3 years old without explicit instruction

Not the kind of stuff you were later taught in “grammar” school

- At least not in English-speaking countries :-)
- Indeed *some* EFL teaching involves something much closer to what we have in mind here

3. Syntax (or Grammar)

Refers to the way words can be arranged in a given language

Grammars (and parsing) are key components in many applications

- Grammar checkers
- Dialogue management
- Question answering
- Information extraction
- Machine translation

4. Syntax, cont'd

There's a useful (traditional) contrast between two perspectives on this topic:

paradigmatic
: What's interchangeable with what?
  - words, phrases, . . .

syntagmatic
: What co-occurs with what?
  - ordering (before/after)
  - marking (prefixes/suffixes)

Key notions that we'll cover

- Categories (paradigmatic)
- Constituency (syntagmatic)
- Heads (syntagmatic)

Key formalism

- Context-free grammars

5. Constituency

Groups of words can be shown to act as single units, called **constituents**

In a given language, these units form coherent classes that behave in similar ways, with respect to

External behavior
: How they relate to other units in the language
  - We can say that in English, noun phrases can come before verbs

Internal structure
: We can describe an internal structure for the class
  - This might involve disjunctions of somewhat unlike sub-classes to do this
  - For example, English noun phrases can consist of a pronoun, a proper noun, or a complex phrase including a common noun

6. Constituency, cont'd: Noun Phrases

We can observe some commonality over the behaviour of the following English phrases:

```
they
Cambodia
my aunt's pen
the reason I can't stay
taking another look at Moby Dick
three french hens
```

One piece of evidence is that they can all precede verbs

- That is, occur in a frame such as

______________ surprised him.

- This is external, paradigmatic evidence

Internal, syntagmatic, evidence would be, for example, to observe that combining

> **determiners** such as "my" or "three"
> **qualifiers** such as "aunt's" or "french"
> **common nouns** such as "pen" or "hens"

usually results in a coherent phrase, which can fit in the above frame

7. Noun phrases in other languages

The internal structure of NPs varies from language to language:

> **English** *these three expensive books* (Dem Num Adj Noun)
> **French** *ces trois livres chers* (Dem Num Noun Adj)
> **Chinese** 這 三 本 昂貴 的 書 (Dem Num clf Adj part Noun)
> **Japanese** これ ら の 三 つ の 高価 な 本 (Dem part pos Num clf pos Adj part Noun)
> **Thai** *hnạngsụu rākhā phæng sām lèm* (Noun Adj Num clf)

8. Grammars and Constituency

There's nothing easy or obvious about how we come up with

- the 'right' set of constituents
- the rules that govern how they combine

That's one of reasons there are so many different theories of grammar and competing analyses of the same data

The approach we'll explore isn't exactly "cutting-edge"

- But it's a good compromise between simplicity and adequacy
- And the technology required to support it is a good introduction to what's needed for most other approaches

9. Context-Free Grammars (CFGs)

Also known as **phrase structure grammars**

- And **Backus-Naur form** is a standardised approach to *notating* CFGs

Making explicit the restrictions on rewriting we started with earlier, a CFG consists of

Terminals
: or **terminal symbols**: words (for now)

Non-terminals
: or **non-terminal symbols**: Names for constituents in a language
  - E.g., **NP** (noun phrase), **VP** (verb phrase), **V** (verb), **S** (sentence)

Rules
: or **productions**, each of which is a pair of

a left-hand side
: a single non-terminal

a right-hand side
: a sequence of any number of terminals and non-terminals

For example:

```
VP → V NP
```

Distinguished symbol

One of the non-terminals

- The starting point for all analyses
- Usually **S**

10. Some preliminary NP Rules

Some overly-simple rules for noun phrases:

```
NP → Det Nominal
NP → ProperNoun
Nominal → Noun | Nominal Noun
```

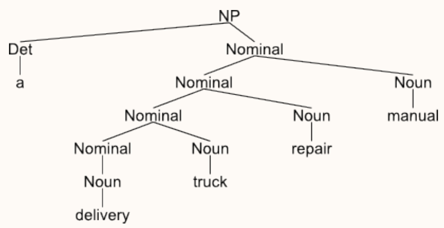
These rules describe two kinds of NPs:

- One that consists of a determiner followed by a nominal
- And another that says that proper names are NPs.

The third rule illustrates:

- A disjunction
  - Two kinds of nominals
  - Not strictly speaking *a* rule
  - Rather a shorthand notation for two rules
- A recursive definition
  - Same non-terminal on the right and left-side of the rule
  - We can see how this works if we consider a noun phrase such as

a delivery truck repair manual



11. A bit more detail on English Grammar

Huddleston and Pullum's *The Cambridge Grammar of the English Language* is 1860 pages long

- !

So we won't cover all of English by a very long way

- Just enough to uncover some key shortcomings of CFGs

We'll look briefly at

- Sentences
- Noun phrases
    - Agreement
- Verb phrases
    - Subcategorisation

12. Sentence Types

Declaratives

A plane left

```
S → NP VP
```

Imperatives

Leave!

```
S → VP
```

Yes-no questions

Did the plane leave?

```
S → Aux NP VP
```

WH questions

When did the plane leave?

```
S → WH-NP Aux NP VP
```

13. Noun Phrases, more carefully

We can identify three quite distinct types of noun phrases:

Pronouns

she, he, we, ...

```
NP -> Pro
```

Proper Nouns

Edinburgh, Star Wars, the Eiffel Tower, ...

```
NP -> PropN
```

Complex noun phrases

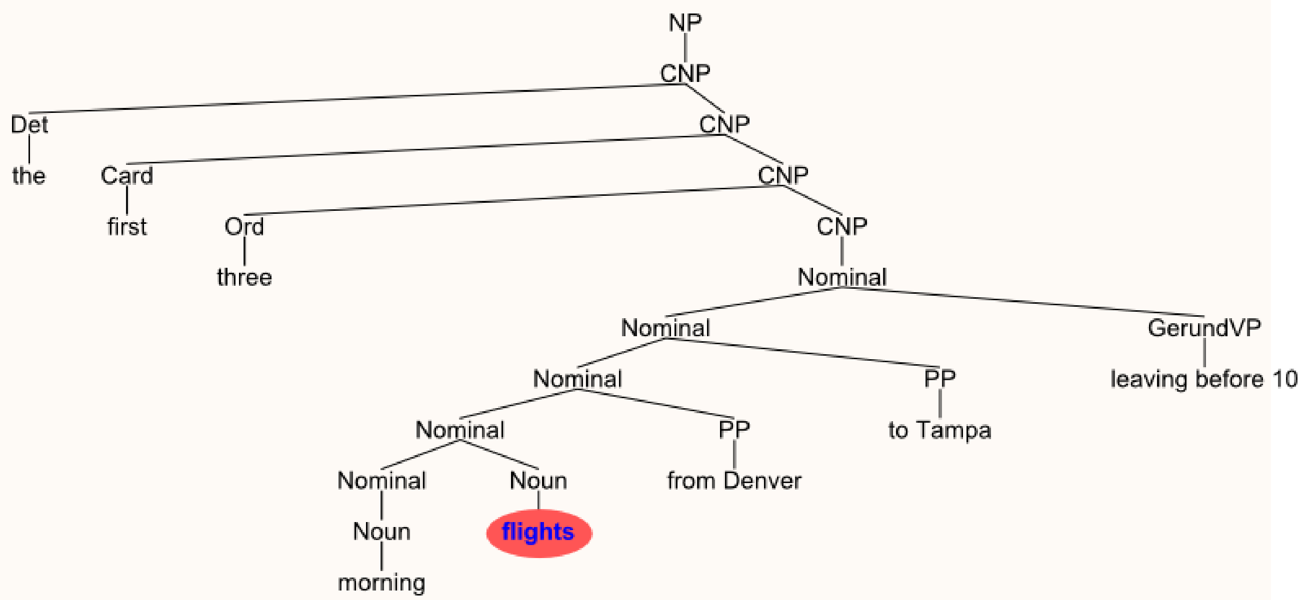
the next prime minister after Thatcher

```
NP -> CNP
```

Consider the following moderately complicated noun phrase:

the first three morning flights from Denver to Tampa leaving before 10

We'll need something along the lines of the following tree



14. NP Structure

That big NP is really about *flights*

- That's its central criticial noun
- Let's call that the **head** of the NP

We can dissect this kind of NP into:

- The constituents that can come *before* the head
- The constituents that can come *after* it

15. Before the nominal: Determiners

Complex noun phrases can start with determiners

```
CNP → Det CNP
```

Determiners can be

Simple lexical items

the, this, a, her

```
Det → Art
```

(Recursive) possessives

simple	Robin's car
complex	Robin's youngest child's toy

```
Det → PropN 's
Det → CNP 's
```

16. Before the nominal: Other premodifiers

Other premodifiers include

- Quantifiers, cardinals, ordinals:

every flight

```
CNP → Quant CNP
```

three flights

```
CNP → Card CNP
```

first flight

```
CNP → Ord CNP
```

- Adjectives and Adjective phrases:

large cars
extremely sleepy baby

```
CNP → AP CNP
AP → Adv AP
AP → Adj
```

There are constraints we haven't captured on the *order* of pre-modifiers:

- Between adjectives and quantifiers:

every eligible candidate
*eligible every candidate

- Between one adjective and another:

big red bus
?red big bus

- Following a common linguistic convention, I'm using an initial asterisk to indicate a word sequence which is *not* in a (natural) language or cannot (should not) be accepted by a formal grammar
- Likewise an initial question mark for a borderline in/out word sequence

17. The nominal: the head and its postmodifiers

Eventually (or even right away), we get to the Nominal

- Including the head, with or without compounding

```
CNP → Nominal
Nominal → Noun
Nominal → Nominal Noun
```

The postmodifiers which stack up behind the head may include

- Prepositional phrases: flight *from Seattle*

- Non-finite clauses (gerundive, infinitive):

flights *arriving before noon*
first flight *to depart*

- Relative clauses:

flights *that serve breakfast*
people *whom the pilot knows*

Similar general (recursive) rules to handle these

```
Nominal → Nominal PP
Nominal → Nominal GerundVP
Nominal → Nominal InfVP
Nominal → Nominal RelClause
```