Context-Free Grammars (CFGs)

L545
Spring 2010

Parsing: Assigning Structure to Sentences

We want to take in an input sentence and assign it a structure.

- **Input:** The man left the room.
- **Output:** (S (NP (DT The) (NN man)) (VP (VBD left) (NP (DT the) (NN room))))

But why this sort of representation?

- Why do we group words as we do?
- Where do we get these categories and what do they mean?

We will later talk about context-free grammars (CFGs) from a formal perspective, but we first want to motivate them linguistically.

Syntax

**Syntax** = the study of the way that sentences are constructed from smaller units.

No “dictionary” for sentences → infinite number of possible sentences.

- The house is large.
- John believes that the house is large.
- Mary says that John believes that the house is large.

There are some basic principles of sentence organization:

- Linear order
- Hierarchical structure (Constituency)
- Subcategorization and Grammatical relations

Constituency tests

There are many “tests” to determine what a constituent is, which are often more like guidelines (i.e., they often work, but they sometimes don’t)

- Preposed/Postposed constructions—i.e., can you move the grouping around?
  1. a. On September seventeenth, I’d like to fly from Atlanta to Denver.
     b. I’d like to fly on September seventeenth from Atlanta to Denver.
     c. I’d like to fly from Atlanta to Denver on September seventeenth.

- Pro-form substitution
  2. John has some very heavy books, but he didn’t want them.
  3. I want to go home, and John wants to do so, too.
Hierarchical structure

Note that constituents appear within other constituents. We can represent this in a bracket form or in a syntactic tree.

Bracket form:

```
[[Many executives] [eat [at [[really fancy] restaurants]]]]
```

Syntactic tree is on the next page ...

Categories

We would also like some way to say that *Many executives* and *really fancy restaurants* are the same type of grouping, or constituent, whereas *at really fancy restaurants* seems to be something else. For this, we will talk about different categories:

- Lexical
- Phrasal

Lexical categories

Lexical categories are simply word classes, or parts of speech. The main ones are:

- verbs: *eat, drink, sleep, ...*
- nouns: *gas, food, lodging, ...*
- adjectives: *quick, happy, brown, ...*
- adverbs: *quickly, happily, well, westward*
- prepositions: *on, in, at, to, into, of, ...*
- determiners/articles: *a, an, the, this, these, some, much, ...*
- conjunctions: *and, but, or, since, while, ...*

Determining lexical categories

How do we determine which category a word belongs to?

- Distribution: where these kinds of words can appear in a sentence.
  
  e.g. Nouns like *mouse* can appear after articles ("determiners") like *the*, while a verb like *eat* cannot.

- Morphology: what kinds of word prefixes/suffixes can a word take?
  
  e.g. Verbs like *walk* can take a *ed* ending to mark them as past tense. A noun like *mouse* cannot.

⇒ Note that if you ever want to do unsupervised learning of POS properties, you'll need these pieces of information.

Closed & Open classes

We can add words to some classes, but not to others.

Open classes: new words can be easily added (tend to carry meaning):

- verbs
- nouns
- adjectives
- adverbs

Closed classes: new words cannot be easily added (tend to be function words):

- prepositions
- determiners
- conjunctions
Phrasal categories

We can also look at the distribution of phrases and see which ones behave in the same way, in order to assign them categories.

- The joggers ran through the park.

What other phrases can we put in place of The joggers?

Susan students you most dogs some children a huge, lovable bear my friends from Brazil the people that we interviewed

Since all of these contain nouns, we consider these to be noun phrases (NPs).

Phrases

Noun Phrases

Noun phrases, like other kinds of phrases, are headed: there is a designated item (the noun) which determines the properties of the whole phrase.

- Before the noun, you can have determiners (and pre-determiners) and adjective phrases.
- After the noun, you can have prepositional phrases, gerunds (and other verbal clauses), and relative clauses.
- You can also have noun-noun compounds.

⇒ General rule: The category of the head word percolates up to the phrase level.

Phrases

Determiner Phrases?

It’s not entirely clear that these phrases should be NPs; maybe they should be DPs.

- There generally must be a noun in an NP, but often there must also be a determiner; in fact, determiners can sometimes appear alone.

(4) *Student/The student* laughed.

(5) *These/These students* think a lot.

- The determiner actually scopes over the noun semantically.

(6) All/Some/No students are happy.

- For some theories, a DP is more uniform with other parts of the syntax.

We’ll still refer to these as NPs, but simply note that it is a non-trivial decision.

Phrases

Verb Phrases: Subcategorization

Verbs tend to drive the analysis of a sentence because they subcategorize for elements.

We can say that verbs have subcategorization frames:

- *sleep*: subject
- *find*: subject, object
- *show*: subject, object, second object
- *want*: subject, object, infinitive verb phrase
- *think*: subject, sentential complement

Phrases

Grammatical relations

Grammatical relations are the basic relations between words in a sentence.

(7) She eats a mammoth breakfast.

- In this sentence, *She* is the subject, while *a mammoth breakfast* is the object.

- In English, the subject must agree in person and number with the verb.
Phrase Structure Rules (PSRs)

We can give rules for building these phrases, i.e., a way to say that a determiner and a noun make up a noun phrase, but a verb and adverb do not.

- Phrase structure rules (PSRs) build larger constituents from smaller ones.
  e.g. S → NP VP
  - A sentence (S) constituent is composed of a noun phrase (NP) constituent and a verb phrase (VP) constituent. (hierarchy)
  - The NP must precede the VP. (linear order)
- Put PSRs together, and you have a context-free grammar (CFG)

Important Properties of Phrase Structure Rules

- recursive = a rule can be reapplied (within its hierarchical structure).
  e.g. NP → NP PP
  PP → P NP
  The property of recursion means that the set of potential sentences in a language is infinite.
- potentially (structurally) ambiguous = have more than one analysis
  (8) I [VP saw [NP [NP the man] [PP with the telescope]]]
  (9) I [VP saw [NP the man] [PP with the telescope]]

Formal definition of CFGs

1. N: a set of non-terminal (phrasal) symbols, e.g., NP, VP, etc.
2. Σ: a set of terminal (lexical) symbols
   - N and Σ are disjoint
3. P: a set of productions (rules) of the form A → α, where A is a non-terminal and α is a collection of terminals and non-terminals
4. S: a designated start symbol

Lingering question: Are CFGs capable of covering language?

Constructions which must be captured by a grammar

- Coordination
- Active & Passive Constructions
- Raising & Control Constructions
- Unbounded Dependency Constructions (UDCs)

Difficulties with coordination

Coordination turns out to have particularly difficult properties for linguistic analysis

- The conjunction of two elements does not obey the same properties as each element.
  (11) a. *Me went to the store.
       b. Me and John went to the store.
- Coordination can be with “unlike” constituents
  (12) Robin is [NP a Republican] and [ADJP proud of it]
- Coordination can be with non-constituents
  (13) John gave me the bread and Mary the sugar.

Coordination

One type of phrase we have not mentioned yet is the coordinate phrase, for example John and Mary

- Coordination can generally apply to any kinds of (identical) phrases
- This makes it ambiguous and cause problems for parsing
  (10) I saw John and Mary left early.

⇒ At some point, a parser has to decide between and joining NPs and joining Ss.
Active & Passive Constructions

It is well-established that sentences occur in both active and passive forms:

(14) a. Sandy saw Kim.
    b. Kim was seen by Sandy.

CFGs can clearly handle such sentences, along the lines of:

• VP → V_{fin} NP
• VP → V_{be} VP_{pass}
• VP_{pass} → V_{pass} (PP_{by})

Relating active and passive constructions

Even if a CFG can license such constructions, questions remain:

• How many rules will it take to capture every relevant grammatical distinction?
• How are the active and passive forms related?
  – Through movement?
  – Through lexical rules?
  – They’re not related?

Raising & Control Constructions

Some verbs look similar in some syntactic contexts, but behave quite differently in others

(15) a. John seems to be happy.
    b. It seems to be raining.
    c. John tries to be happy.
    d. *It tries to be raining.

Generalization:

• Raising verbs (e.g., seem): the subject of the higher clause is the “same” as the subject of the lower clause
• Control (or equi) verbs (e.g., try): the subject of the higher clause “controls” the subject of the lower clause, but has certain restrictions on it.

Capturing the raising/control generalizations

How do we distinguish raising and control verbs in CFGs?

• In both cases, it seems like we have the pattern NP V V_{inf}

Solutions seem to require one or more of the following:

• An empty subject in the lower clause.
• Sharing of subjects (or subject properties) between upper and lower verbs, perhaps involving the addition of new features.
• A closer connection to sentence semantics.

Unbounded Dependency Constructions (UDCs)

An unbounded dependency construction has an element realized non-locally and:

• involves constituents with different functions
• involves constituents of different categories
• is in principle unbounded

Example: Wh-elements

Wh-elements can have different functions:

(16) a. Who did Hobbs see _?
    b. Who do you think _ saw the man?
    c. Who did Hobbs give the book to _?
    d. Who did Hobbs consider _ to be a fool?

Wh-elements can also occur in subordinate clauses:

(17) a. I asked who the man saw _.
    b. I asked who the man considered _ to be a fool.
    c. I asked who Hobbs gave the book to _.
    d. I asked who you thought _ saw Hobbs.
Different categories can be extracted:

(18) a. Which man did you talk to _ ?
   b. [To [which man]] did you talk _ ?
   c. [How ill] has the man been _ ?
   d. [How frequently] did you see the man _ ?

This sometimes provides multiple options for a constituent:

(19) a. Who does he rely [on _ ]?
   b. [On whom] does he rely _ ?

Unboundedness:

(20) a. Who do you think Hobbs saw _ ?
   b. Who do you think Hobbs said he saw _ ?
   c. Who do you think Hobbs said he imagined that he saw _ ?

Accounting for UDCs

How does one account for UDCs?

- Invoke a notion of movement during an analysis
- Include features which “pass” information about the non-local element
- Use some formalism more powerful than a CFG (e.g., Tree-Adjoining Grammar)