CKY Parsing

Ling571
Deep Processing Approaches to NLP
January 11, 2016
Roadmap

- Motivation:
  - Inefficiencies of parsing-as-search

- Strategy: Dynamic Programming

- Chomsky Normal Form
  - Weak and strong equivalence

- CKY parsing algorithm
Top-down parsing (DFS)
Bottom-Up Parsing

- Try to find all trees that span the input
  - Start with input string
    - Book that flight.

- Use all productions with current subtree(s) on RHS
  - E.g., N → Book; V → Book

- Stop when spanned by S (or no more rules apply)
Bottom-Up Search

```
VP
  /   \\   /
Verb Det Noun
  Book that flight
NP
Nominal
```
Bottom-Up Search
Pros and Cons of Bottom-Up Search

- **Pros:**
  - Will not explore trees that don’t match input
  - Recursive rules less problematic
  - Useful for incremental/fragment parsing

- **Cons:**
  - Explore subtrees that will not fit full sentences
Parsing Challenges

- Ambiguity
- Repeated substructure
- Recursion
Parsing Ambiguity

- Many sources of parse ambiguity
  - Lexical ambiguity
    - Book/N; Book/V

- Structural ambiguity: Main types:
  - Attachment ambiguity
    - Constituent can attach in multiple places
      - *I shot an elephant in my pyjamas.*

- Coordination ambiguity
  - Different constituents can be conjoined
    - *Old men and women*
Ambiguity

Speech and Language Processing - Jurafsky and Martin
Disambiguation

- **Global ambiguity:**
  - Multiple complete alternative parses
  - Need strategy to select correct one
    - Approaches exploit other information
      - Statistical
        - Some prepositional structs more likely to attach high/low
        - Some phrases more likely, e.g., (old (men and women))
      - Semantic
      - Pragmatic
        - E.g., elephants and pyjamas
  - Alternatively, keep all

- **Local ambiguity:**
  - Ambiguity in subtree, resolved globally
Repeated Work

- Top-down and bottom-up parsing both lead to repeated substructures
  - Globally bad parses can construct good subtrees
    - But overall parse will fail
    - Require reconstruction on other branch

- No static backtracking strategy can avoid

- Efficient parsing techniques require storage of shared substructure
  - Typically with dynamic programming

- Example: *a flight from Indianapolis to Houston on TWA*
Shared Sub-Problems

NP
  └── Det
     └── a
  │    └── Nominal
  │         └── PP
  │             └── Noun
  │                     └── from Indianapolis...
  │                                └── flight
Shared Sub-Problems
Shared Sub-Problems

```
  Det
   | a
  NP  Nominal
     Nominal  PP
     Nominal  PP
      Nominal  to Houston
       Noun  PP
          Nominal  PP
           Noun  PP
             Nominal  PP
              Noun  PP
                Nominal  PP
                  Noun  PP
                    Noun
                      Nominal
                        Nominal
                          Nominal
                            Nominal
                              Nominal
                                Nominal
                                  Nominal
                                    Nominal
                                      Nominal
                                        Nominal
                                          Nominal
                                            Nominal
                                              Nominal
                                                Nominal
                                                  Nominal
                                                    Nominal
                                                      Nominal
                                                        Nominal
                                                          Nominal
                                                            Nominal
                                                              Nominal
                                                                Nominal
                                                                  Nominal
                                                                    Nominal
                                                                      Nominal
                                                                        Nominal
                                                                          Nominal
                                                                            Nominal
                                                                             Nominal
                                                                              Nominal
                                                                                Nominal
                                                                                  Nominal
                                                                                      Nominal
                                                                                       Nominal
                                                                                         Nominal
                                                                                           Nominal
                                                                                             Nominal
                                                                                                 Nominal
                                                                                                   Nominal
                                                                                                     Nominal
                                                                                                       Nominal
                                                                                                         Nominal
                                                                                                          Nominal
                                                                                                               Nominal
                                                                                                                Nominal
                                                                                                                  Nominal
                                                                                                                    Nominal
                                                                                                                     Nominal
                                                                                                                      Nominal
                                                                                                                        Nominal
                                                                                                                             Nominal
                                                                                                                               Nominal
                                                                                                                                  Nominal
                                                                                                                                   Nominal
                                                                                                                                    Nominal
                                                                                                                                       Nominal
                                                                                                                                            Nominal
                                                                                                                                                Nominal
                                                                                                                                                    Nominal
                                                                                                                                                       Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                                    Nominal
                                                                                                                                            `
Recursion

- Many grammars have recursive rules
  - E.g., $S \rightarrow S \text{ Conj } S$

- In search approaches, recursion is problematic
  - Can yield infinite searches
    - Esp., top-down
Dynamic Programming

- Challenge: Repeated substructure $\rightarrow$ Repeated work

- Insight:
  - Global parse composed of parse substructures
  - Can record parses of substructures

- Dynamic programming avoids repeated work by tabulating solutions to subproblems
  - Here, stores subtrees
Parsing w/Dynamic Programming

- Avoids repeated work
- Allows implementation of (relatively) efficient parsing algorithms
  - Polynomial time in input length
    - Typically cubic ($n^3$) or less
- Several different implementations
  - Cocke-Kasami-Younger (CKY) algorithm
  - Earley algorithm
  - Chart parsing
Chomsky Normal Form (CNF)

- CKY parsing requires grammars in CNF

- Chomsky Normal Form
  - All productions of the form:
    - $A \rightarrow B \ C$, or
    - $A \rightarrow a$

- However, most of our grammars are not of this form
  - E.g., $S \rightarrow \text{Wh-NP} \ \text{Aux} \ \text{NP} \ \text{VP}$

- Need a general conversion procedure
  - Any arbitrary grammar can be converted to CNF
Grammar Equivalence and Form

- Grammar equivalence
  - Weak: Accept the same language, May produce different analyses
  - Strong: Accept same language, Produce same structure
CNF Conversion

- Three main conditions:
  - Hybrid rules:
    - INF-VP $\rightarrow$ to VP
  - Unit productions:
    - $A \rightarrow B$
  - Long productions:
    - $A \rightarrow B \; C \; D$
CNF Conversion

- Hybrid rule conversion:
  - Replace all terminals with dummy non-terminals
  - E.g., INF-VP → to VP
    - INF-VP → TO VP; TO → to

- Unit productions:
  - Rewrite RHS with RHS of all derivable non-unit productions
    - If $A \Rightarrow^* B$ and $B \Rightarrow w$, then add $A \Rightarrow w$
CNF Conversion

- Long productions:
  - Introduce new non-terminals and spread over rules
  - $S \rightarrow$ Aux NP VP
    - $S \rightarrow$ X1 VP; $X1 \rightarrow$ Aux NP

- For all non-conforming rules,
  - Convert terminals to dummy non-terminals
  - Convert unit productions
  - Binarize all resulting rules
<table>
<thead>
<tr>
<th><strong>Grammar</strong></th>
<th><strong>in CNF</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>S → NP VP</td>
<td>S → NP VP</td>
</tr>
<tr>
<td>S → Aux NP VP</td>
<td>S → X1 VP</td>
</tr>
<tr>
<td>X1 → Aux NP</td>
<td></td>
</tr>
<tr>
<td>S → VP</td>
<td>S → book</td>
</tr>
<tr>
<td></td>
<td>S → Verb NP</td>
</tr>
<tr>
<td></td>
<td>S → X2 PP</td>
</tr>
<tr>
<td></td>
<td>S → Verb PP</td>
</tr>
<tr>
<td></td>
<td>S → VP PP</td>
</tr>
<tr>
<td>NP → Pronoun</td>
<td>NP → I</td>
</tr>
<tr>
<td>NP → Proper-Noun</td>
<td>NP → TWA</td>
</tr>
<tr>
<td>NP → Det Nominal</td>
<td>NP → Det Nominal</td>
</tr>
<tr>
<td>Nominal → Noun</td>
<td>Nominal → book</td>
</tr>
<tr>
<td>Nominal → Nominal Noun</td>
<td>Nominal → Nominal Noun</td>
</tr>
<tr>
<td>Nominal → Nominal PP</td>
<td>Nominal → Nominal PP</td>
</tr>
<tr>
<td>VP → Verb</td>
<td>VP → book</td>
</tr>
<tr>
<td>VP → Verb NP</td>
<td>VP → Verb NP</td>
</tr>
<tr>
<td>VP → Verb NP PP</td>
<td>VP → X2 PP</td>
</tr>
<tr>
<td></td>
<td>X2 → Verb NP</td>
</tr>
<tr>
<td>VP → Verb PP</td>
<td>VP → Verb PP</td>
</tr>
<tr>
<td>VP → VP PP</td>
<td>VP → VP PP</td>
</tr>
<tr>
<td>PP → Preposition NP</td>
<td>PP → Preposition NP</td>
</tr>
</tbody>
</table>
CKY Parsing

- Cocke-Kasami-Younger parsing algorithm:
  - (Relatively) efficient bottom-up parsing algorithm based on tabulating substring parses to avoid repeated work

- Approach:
  - Use a CNF grammar
  - Build an \((n+1) \times (n+1)\) matrix to store subtrees
    - Upper triangular portion
  - Incrementally build parse spanning whole input string
Dynamic Programming in CKY

- Key idea:
  - For a parse spanning substring \([i,j]\), there exists some \(k\) such there are parses spanning \([i,k]\) and \([k,j]\)
    - We can construct parses for whole sentence by building up from these stored partial parses

- So,
  - To have a rule \(A \rightarrow B C\) in \([i,j]\),
    - We must have \(B\) in \([i,k]\) and \(C\) in \([k,j]\), for some \(i < k < j\)
      - CNF grammar forces this for all \(j > i + 1\)
CKY

- Given an input string $S$ of length $n$,
  - Build table $(n+1) \times (n+1)$
  - Indexes correspond to inter-word positions
    - E.g., 0 Book 1 That 2 Flight 3
  - Cells $[i,j]$ contain sets of non-terminals of ALL constituents spanning $i,j$
    - $[j-1,j]$ contains pre-terminals
    - If $[0,n]$ contains Start, the input is recognized
function CKY-PARSE(words, grammar) returns table

for j ← from 1 to LENGTH(words) do
    table[j - 1, j] ← \{A \mid A \to \text{words}[j] \in \text{grammar}\}
    for i ← from j - 2 downto 0 do
        for k ← i + 1 to j - 1 do
            table[i, j] ← table[i, j] ∪
            {A \mid A \to BC \in \text{grammar},
             B \in \text{table}[i, k],
             C \in \text{table}[k, j]}
Is this a parser?
CKY Parsing

- Table fills:
  - Column-by-column
  - Left-to-right
  - Bottom-to-top

- Why?
  - Necessary info available (below and left)
  - Allows online sentence analysis
    - Works across input string as it arrives
<table>
<thead>
<tr>
<th>Book</th>
<th>the</th>
<th>flight</th>
<th>through</th>
<th>Houston</th>
</tr>
</thead>
<tbody>
<tr>
<td>S, VP, Verb</td>
<td>S, VP, X2</td>
<td>S, VP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[0, 1]</td>
<td>[0, 3]</td>
<td>[0, 4]</td>
<td>[0, 5]</td>
<td></td>
</tr>
<tr>
<td>Dot</td>
<td>NP</td>
<td>Nominal, Noun</td>
<td>Nominal</td>
<td></td>
</tr>
<tr>
<td>[1, 2]</td>
<td>[1, 3]</td>
<td>[1, 4]</td>
<td>[1, 5]</td>
<td></td>
</tr>
<tr>
<td>Nominal, Noun</td>
<td>Prep</td>
<td>PP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2, 3]</td>
<td>[2, 4]</td>
<td>[2, 5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prep</td>
<td>NP, Proper-Noun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3, 4]</td>
<td>[3, 5]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Book the flight through Houston
Filling CKY cell
<table>
<thead>
<tr>
<th>Book</th>
<th>the</th>
<th>Flight</th>
<th>Through</th>
<th>Houston</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN, VB, Nominal, VP, S [0,1]</td>
<td>[0,2]</td>
<td>S, VP, X2 [0,3]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Det [1,2]</td>
<td>NP [1,3]</td>
<td>NN, Nominal [2,3]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>