UPPSALA
UNIVERSITET

Earley's algorithm
Discussion

Syntactic analysis/parsing

2024

Sara Stymne
Department of Linguistics and Philology

Based on slides by Marco Kuhlmann

UPPSALA UNIVERSITET

Today

- Presentation of Earley's algorithm
- Working through an exercise on it
- If time:
  - Advanced PCFG

UPPSALA
UNIVERSITET

The Earley algorithm

CKY versus Earley

- The CKY algorithm has two disadvantages:
  - It can only handle restricted grammars.
  - It does not use top–down information.
- The Earley algorithm does not have these:
  - It can handle arbitrary grammars.
  - Is does use top–down information.
  - On the downside, it is more complicated.

The algorithm

- Start with the start symbol *S*.
- Take the leftmost nonterminal and predict all possible expansions.
- If the next symbol in the expansion is a word, match it against the input sentence (scan); otherwise, repeat.
- If there is nothing more to expand, the subtree is complete; in this case, continue with the next incomplete subtree.

Dotted rules

- A dotted rule is a partially processed rule.

  Example: S → NP • VP

- The dot can be placed in front of the first symbol, behind the last symbol, or between two symbols on the right-hand side of a rule.
- The general form of a dotted rule thus is $A \rightarrow \alpha \bullet \beta$, where $A \rightarrow \alpha\beta$ is the original, non-dotted rule.

Inference rules

Axiom

$$[0, 0, S \rightarrow \bullet\ \alpha]$$

$S \rightarrow \alpha$

Predict

$$\frac{[i, j, A \rightarrow \alpha \bullet B\ \beta]}{[j, j, B \rightarrow \bullet\ \gamma]}$$

$B \rightarrow \gamma$

Scan

$$\frac{[i, j, A \rightarrow \alpha \bullet a\ \beta]}{[i, j+1, A \rightarrow \alpha\ a \bullet \beta]}$$

$w_j = a$

Complete

$$\frac{[i, j, A \rightarrow \alpha \bullet B\ \beta] \quad [j, k, B \rightarrow \gamma \bullet]}{[i, k, A \rightarrow \alpha\ B \bullet \beta]}$$

Pseudo code 1

```
function EARLEY-PARSE(words, grammar) returns chart

  ENQUEUE((γ → • S, [0,0]), chart[0])
  for i ← from 0 to LENGTH(words) do
    for each state in chart[i] do
      if INCOMPLETE?(state) and
              NEXT-CAT(state) is not a part of speech then
          PREDICTOR(state)
      elseif INCOMPLETE?(state) and
              NEXT-CAT(state) is a part of speech then
          SCANNER(state)
      else
          COMPLETER(state)
    end
  end
  return(chart)
```

Image from Jurafsky & Martin, edition 2

Pseudo code 2

procedure PREDICTOR($(A \rightarrow \alpha \bullet B\ \beta,\ [i,j])$)
  for each $(B \rightarrow \gamma)$ **in** GRAMMAR-RULES-FOR($B, grammar$) **do**
    ENQUEUE($(B \rightarrow \bullet\ \gamma,\ [j,j]), chart[j]$)
  end

procedure SCANNER($(A \rightarrow \alpha \bullet B\ \beta,\ [i,j])$)
  if B $\subset$ PARTS-OF-SPEECH($word[j]$) **then**
    ENQUEUE($(B \rightarrow word[j],\ [j,j+1]), chart[j+1]$)

procedure COMPLETER($(B \rightarrow \gamma \bullet,\ [j,k])$)
  for each $(A \rightarrow \alpha \bullet B\ \beta,\ [i,j])$ **in** $chart[j]$ **do**
    ENQUEUE($(A \rightarrow \alpha\ B \bullet \beta,\ [i,k]), chart[k]$)
  end

Image from Jurafsky & Martin, edition 2

Recogniser/parser

- When parsing is complete, is there a chart entry?
  [0, n, S → α •]
  - Recognizer
- If we want a parser, we have to add back pointers, and retrieve a tree
- Earley's algorithm can be used for PCFGs, but it is more complicated than for CKY

UPPSALA UNIVERSITET

Coming up

- Monday, Feb. 26: Supervision
  - Assignment 3
- Wednesday, Feb 28: Lecture graph-based parsing
- Monday, March 4: Seminar 2

UPPSALA UNIVERSITET

Coming deadlines

- Project:
  - Choose individual or pair project
    - Sign up in Studium
      - Decide on your topic: information on the web page!
      - Proposal: February 26
- Project seminar moved to March 25, 9-12!
- Assignment 2: Feb. 22
- Assignment 3: March 11 (moved from March 4)

UPPSALA UNIVERSITET

Earley's algorithm as a project

- Suitable scope: implement a recognizer
  - Use a small toy grammar to show that it works
  - A parser is a bit too much to code, but you are expected to discuss it in your report
  - Also: read at least one relevant article (the math can be a bit challenging, so OK to read it at a higher level of abstraction)

UPPSALA UNIVERSITET

Missing a deadline

- If you miss a deadline (and you have not contacted Sara beforehand due to extraordinary circumstances)
- You can submit at a second deadline:
  - All assignments: April 1
  - Project and seminar report: April 15