Guaranteed Parsing

The CYK algorithm for parsing works with any context-free language.

CYK

The CYK algorithm is named after Cocke, Younger and Kasami. It assumes CFG G in Chomsky Normal Form.

It uses dynamic programming. . .

Generalized Problem

Say input $w = w_1 w_2 \dots w_n$. Then, let $w_{i,j}$ be substring $w_i w_{i+1} \dots w_j$. The problem one solves is:

which variables produce which substrings.

A Recursive Formula

In general, suppose we want to know if variable $A \overset{*}{\Longrightarrow} w_{i,j}$ where $|w_{i,j}| \geq 2$:

The first step in derivation must be production of form $A \rightarrow EF$. So, $w_{i,j}$ can be split into two pieces: the first generated from E and the second from F. (But we don't know where split occurs.) Hence the recursive formula:

> *Consider all productions* $A \rightarrow EF$. *For all possible* k *from* i *up to* $j-1$, *ask whether* $E \overset{*}{\Longrightarrow} w_{i,k}$ *and* $F \overset{*}{\Longrightarrow} w_{k+1,j}$.

The Overall Algorithm

To make efficient, answer question for smaller strings first, keeping results in a table.

CYK algorithm. *1. Start by answering for each i and each variable A whether $A \overset{*}{\Longrightarrow} w_{i,i}$. (Look at unit productions.)*

2. Then answer for each i and each variable A whether $A \overset{*}{\Longrightarrow} w_{i,i+1}$. (Use recursive formula.)

3. Repeat for all $w_{i,i+2}$, then all $w_{i,i+3}$, and so on.

Eventually, we determine variables for $w = w_{1,n}$.

Example Parsing

Consider CFG with start variable S:

$$
\begin{aligned}
S &\rightarrow ST \mid TU \mid \mathsf{b} \\
T &\rightarrow SU \mid \mathsf{a} \\
U &\rightarrow SS \mid \mathsf{b}
\end{aligned}
$$

Consider input string $w = \mathsf{aababb}$.

Example Table for aababb

		finish					
		1	2	3	4	5	6
start	1	T	.	.	.	S	S, T, U
	2		T	S	S	S, T, U	S, T, U
	3			S, U	S	T, U	S, T, U
	4				T	S	S, T, U
	5					S, U	T, U
	6						S, U

For example, entry in row 3 column 5 says that variables T and U generate $w_{3,5}$: T is here since $T \rightarrow SU$ and $S \overset{*}{\Longrightarrow} w_{3,4}$, $U \overset{*}{\Longrightarrow} w_{5,5}$.

Practice

For the earlier grammar

1: $S \rightarrow \texttt{r}L$
2: $L \rightarrow L\texttt{,}I$
3: $L \rightarrow I$
4: $I \rightarrow \texttt{v}$

Convert to Chomsky Normal Form, and then apply the CYK algorithm to the string `rv,v,v`.

Solution to Practice

$S \to RL$

$L \to LF \mid \mathsf{v}$

$F \to CI$

$I \to \mathsf{v}$

$C \to \mathsf{,}$

$R \to \mathsf{r}$

	1	2	3	4	5	6
1	R	S	.	S	.	S
2		L, I	.	L	.	L
3			C	F	.	.
4				L, I	.	L
5					C	F
6						L, I

Summary

The CYK algorithm can be used to parse any context-free language.