Split Head Automaton Grammars

These slides give a refomulation of material in "Efficient Parsing for Bilexical Context Free Grammars and Head Automata Grammars" by Jason Eisner and Giorgia Satta in ACL-99.

Let V be a set of words and let g and t be positive integers. A V-g-t-split head automaton grammar is a set of productions of the following forms where $w,\ y \in V$, $i \in [1,\ g]$, and $n, m \in [1,\ t]$.

$$
\begin{array}{lcl}
S & \rightarrow & X_w \\
X_w & \rightarrow & L_w^i w R_w^i
\end{array}
$$

$$
\begin{array}{lcl}
L_w^i & \rightarrow & L_w^{i,n} \\
L_w^{i,n} & \rightarrow & X_y L_w^{i,m} \\
L_w^{i,1} & \rightarrow & \epsilon
\end{array}
$$

$$
\begin{array}{lcl}
R_w^i & \rightarrow & R_w^{i,n} \\
R_w^{i,n} & \rightarrow & R_w^{i,m} X_y \\
R_w^{i,1} & \rightarrow & \epsilon
\end{array}
$$

The grammar will be called deterministic if for all values w, i, m and y there is at most one production of the form $R_w^{i,n} \rightarrow R_w^{i,m} X_y$ and at most one production of the form $L_w^{i,n} \rightarrow X_y L_w^{i,m}$. If we restrict uses of the two epsilon productions to cases where w is the appropriate adjacent word, then running the CKY algorithm on a string s takes $O(|s|^4 gt)$ time for deterministic grammars — the dominating production is $R_w^{i,n} \rightarrow R_w^{i,m} X_y$ and its dual.

A Grammar Transformation

Replace each production $R_w^{i,n} \rightarrow R_w^{i,m} X_y$ by the following pair of productions for each sense p of y. Here $[R_w^{i,n}/R_y^p]$ represents a nonterminal symbol.

$$[R_w^{i,n}/R_y^p] \rightarrow R_w^{i,m}\ L_y^p\ y$$

$$R_w^{i,n} \rightarrow [R_w^{i,n}/R_y^p]\ R_y^p$$

Perform the dual replacement for productions of the form $L_w^{i,n} \rightarrow X_y L_w^{i,m}$. Then remove all productions of the form $X_w \rightarrow R_w^i w L_w^i$ and replace each production $S \rightarrow X_w$ with the set of productions of the form $S \rightarrow R_w^i w L_w^i$. Note that we have now removed all occurances of nonterminals of the form X_w. Also note we have kept productions of the form $R_w^{i,1} \rightarrow \epsilon$ and $R_w^i \rightarrow R_w^{i,n}$ and their duals.

Under the appropriate restriction on the use of ϵ productions, CKY parsing maintains the follwing invariants plus their duals.

- If the chart contains $w_{h,k} \in R_u^{i,n}$ or $w_{h,k} \in R_u^i$ then $u = w_{h-1}$.
- If the chart contains $w_{h,k} \in [R_u^{i,n}/R_v^p]$ then $u = w_{h-1}$ and $v = w_k$.

These invariants imply that running CKY on a given string s takes $O(|s|^3 g^2 t)$ time for deterministic grammars.