Applications of Tree Automata
(in memoriam of Ferenc Gécseg)

Andreas Maletti

Institute for Natural Language Processing
Universität Stuttgart, Germany

maletti@ims.uni-stuttgart.de

Szeged — October 9, 2015
Tree Automata
THATCHER & WRIGHT (1968):  
*Generalized Finite Automata Theory with an Application to a Decision Problem of Second-Order Logic.*  

ROUNDS (1970):  
*Mappings and Grammars on Trees.*  

GÉCSEG (1977):  
*Universal Algebras and Tree Automata.*  
Proc. FCT: 98–112
tree automaton: \((Q, \Sigma, I, R)\)

- finite set \(Q\) of states
- finite set \(\Sigma\) of terminals
- initial states \(I \subseteq Q\)
- finite set \(R\) of rules of the form \(q \rightarrow \sigma(q_1, \ldots, q_k)\)
  
  \((\sigma \in \Sigma, k \geq 0, q, q_1, \ldots, q_k \in Q)\)

element rules:

\[
\begin{align*}
q_4 & \rightarrow q_5 \quad q_2 \quad q_3 \\
q_0 & \rightarrow q_1 \quad q_4 \\
q_0 & \rightarrow q_6 \quad q_2
\end{align*}
\]
derivation semantics: $\xi \Rightarrow \zeta$ for sentential forms $\xi, \zeta \in T_{\Sigma}(Q)$ if there exist leaf position $w$ in $\xi$ and rule $q \rightarrow r$ in $R$

$$\xi = \xi[q]_w \quad \zeta = \xi[r]_w$$

recognized tree language:

$$\{ t \in T_{\Sigma} \mid \exists q_0 \in I: q_0 \Rightarrow^* t \}$$
example derivation:

\[
q_0 \Rightarrow S \xrightarrow{q_1 \ q_4} S \xrightarrow{q_1 \ VP \ q_2 \ q_3}
\]

example rules:

\[
q_4 \rightarrow VP \xrightarrow{q_5 \ q_2 \ q_3}
\]

\[
q_0 \rightarrow S \xrightarrow{q_1 \ q_4}
\]

\[
q_0 \rightarrow S \xrightarrow{q_6 \ q_2}
\]
Tree Automaton

Standard references

GÉCSEG & STEINBY (1984):
Tree Automata.
Akadémiai Kiadó


GÉCSEG & STEINBY (1997):
Tree Languages.
In Handbook of Formal Languages 3: 1–68, Springer
Parsing
Parsing: determining the syntactic structure of a sentence

Example: We must bear in mind the Community as a whole
Parsing: determining the syntactic structure of a sentence

Example: We must bear in mind the Community as a whole

POS-tag: part-of-speech tag, “class” of a word
short history:

- **before 1990**
  - hand-crafted CFG
  - corrections and selection by human annotators

- **1990s**
  - PENN tree bank (1M words)
  - weighted CFG
  - WALL STREET JOURNAL tree bank (30M words)

- **since 2000**
  - weighted tree automata
popular parsers (for English):

- **COLLINS parser**: CFG with manual subcategorization
  
  [COLLINS, 1999]

- **STANFORD parser**: CFG with manual subcategorization
  
  [KLEIN, MANNING, 2003]

- **BLLIP parser**: CFG with manual subcategorization and reranking
  
  [CHARNIAK, JOHNSON, 2005]

- **BERKELEY parser**: CFG with automatic subcategorization
  
  [PETROV, KLEIN, 2007]
tags:

- official tags often conservative
  - English: \(\approx 50\) tags
  - German: \(\gg 200\) tags

ADJA-Sup-Dat-Sg-Fem
tags:

- official tags often conservative
  - English: ≈ 50 tags
  - German: ≫ 200 tags

- modern parsers use refined tags → subcategorization
tags:

- official tags often conservative
  - English: ≈ 50 tags
  - German: ≫ 200 tags
- modern parsers use refined tags → subcategorization
- but return parse over official tags → relabeling

ADJA-Sup-Dat-Sg-Fem
read off CFG productions:

\[
\begin{align*}
S & \rightarrow \text{NP } \text{VP} \\
\text{PRP$} & \rightarrow \text{My} \\
\text{VP} & \rightarrow \text{VBZ} \\
\text{NP} & \rightarrow \text{PRP} \\
\text{VP} & \rightarrow \text{VBD } \text{ADVP} \\
\text{ADVP} & \rightarrow \text{RB}
\end{align*}
\]

\[
\begin{align*}
\text{NP} & \rightarrow \text{PRP$ } \text{NN} \\
\text{NN} & \rightarrow \text{dog} \\
\text{VBZ} & \rightarrow \text{sleeps} \\
\text{PRP} & \rightarrow \text{l} \\
\text{VBD} & \rightarrow \text{scored} \\
\text{RB} & \rightarrow \text{well}
\end{align*}
\]
Weights for disambiguation

\[
\begin{align*}
S & \rightarrow NP \ v_{\text{We}} \ v_{\text{saw}} \ NP \\
S & \rightarrow NP \ v_{\text{We}} \ v_{\text{saw}} \ NP \\
S & \rightarrow NP \ v_{\text{We}} \ v_{\text{saw}} \ NP \\
S & \rightarrow NP \ v_{\text{We}} \ v_{\text{saw}} \ NP
\end{align*}
\]
BERKELEY parser [Reference]:

BLLIP parser:
State-of-the-art models

- CFG with subcategorization \((\text{CFG}_{\text{sub}})\)
- Tree substitution grammars with subcategorization \((\text{TSG}_{\text{sub}})\) \([\text{SHINDO et al., 2012}]\)

(both as expressive as weighted tree automata)

- Other models
State-of-the-art models

- CFG with subcategorization \((\text{CFG}_{\text{sub}})\)
- Tree substitution grammars with subcategorization \((\text{TSG}_{\text{sub}})\) [Shindo et al., 2012]

(both as expressive as weighted tree automata)

- Other models

Experiment [Shindo et al., 2012]

<table>
<thead>
<tr>
<th>grammar model</th>
<th>(F_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wCFG</td>
<td>72.6</td>
</tr>
<tr>
<td>wTSG ([\text{Cohn et al., 2010}])</td>
<td>84.7</td>
</tr>
<tr>
<td>wCFG_{sub} ([\text{Petrov, 2010}])</td>
<td>91.8</td>
</tr>
<tr>
<td>wTSG_{sub} ([\text{Shindo et al., 2012}])</td>
<td>92.4</td>
</tr>
</tbody>
</table>
Parsing

grammar with subcategorization:
- a grammar $G$ generating $L(G) \subseteq T_\Sigma(W)$
- a (total) mapping $\rho: \Sigma \rightarrow \Delta$

(grammar with relabeling) (subcategorized trees) (functional relabeling)
grammar with subcategorization:
- a grammar $G$ generating $L(G) \subseteq T_\Sigma(W)$
- a (total) mapping $\rho : \Sigma \rightarrow \Delta$

(grammar with relabeling) (subcategorized trees) (functional relabeling)
grammar with subcategorization:

- a grammar $G$ generating $L(G) \subseteq T_\Sigma(W)$
- a (total) mapping $\rho: \Sigma \rightarrow \Delta$

(grammar with relabeling) (subcategorized trees) (functional relabeling)

$S$

$NP$

$PRP$

My

$NN$

dog

$VP$

$VBZ$

sleeps

Semantics

$L(G, \rho) = \rho(L(G)) = \{\rho(t) | t \in L(G)\}$

Language class: REL ($L$) for language class $L$
grammar with subcategorization:
- a grammar $G$ generating $L(G) \subseteq T_\Sigma(W)$
- a (total) mapping $\rho: \Sigma \rightarrow \Delta$

(grammar with relabeling) (subcategorized trees) (functional relabeling)

Semantics

$$L(G, \rho) = \rho(L(G)) = \{\rho(t) \mid t \in L(G)\}$$

Language class: $\text{REL}(\mathcal{L})$ for language class $\mathcal{L}$
Theorem

\[ \text{REL}(w_{\text{CFL}}) = \text{REL}(w_{\text{TSL}}) = w_{\text{RTL}} \]
Theorem

\[ \text{REL}(w_{\text{CFL}}) = \text{REL}(w_{\text{TSL}}) = w_{\text{RTL}} \]
Theorem

\[ \text{REL}(w\text{CFL}) = \text{REL}(w\text{TSL}) = w\text{RTL} \]

hence: subcategorization \( \approx \) finite-state
Comparison:

1. rule of subcategorized grammar:

   \[ S-1 \rightarrow \text{ADJP-2} \ S-1 \quad \text{weight: 0.003545} \]

   with relabeling \( \rho(S-1) = S, \ldots \)
Comparison:

1. rule of subcategorized grammar:

\[ S-1 \rightarrow ADJP-2 \ S-1 \quad \text{weight: 0.003545} \]

with relabeling \( \rho(S-1) = S, \ldots \)

2. corresponding rule of tree automaton

\[ S-1 \rightarrow S(ADJP-2, \ S-1) \quad \text{weight: 0.003545} \]
XML
History:

- **XML** = eXtensible Markup Language
- developed by W3C
- readable by humans and machines
- open standard; widely supported
- uncountable applications (MS Office, Apple iWork, etc.)

Example:

```xml
<UL>
  <LI> first item </LI>
  <LI> second item </LI>
  ...
  <LI> last item </LI>
</UL>
```
Similar development:

- **2000:**
  - original DTD (Document Type Definition) specification
  - essentially a CFG

- **2005–2010:**
  - more expressive specifications
  - DSD, XML-Schema, ISO Relax NG, etc.
  - all implementable by (unranked) tree automata

- **since 2010:**
  - DTD largely irrelevant
Similar development:

- **2000:**
  - original DTD (Document Type Definition) specification
  - essentially a CFG

- **2005–2010:**
  - more expressive specifications
  - DSD, XML-Schema, ISO Relax NG, etc.
  - all implementable by (unranked) tree automata

- **since 2010:**
  - DTD largely irrelevant

\[
\text{tree automata as basis for XML documents}
\]
Selected Literature

KLEIN, MANNING: Accurate Unlexicalized Parsing
Proc. ACL 2003

MURATA, LEE, MANI, KAWAGUCHI: Taxonomy of XML schema languages using formal language theory
ACM Trans. Internet Technology 2005

PETROV, BARRETT, THIBAUX AND KLEIN: Learning Accurate, Compact, and Interpretable Tree Annotation. Proc. ACL 2006

SHINDO, MIYAO, FUJINO AND NAGATA: Bayesian Symbol-Refined Tree Substitution Grammars for Syntactic Parsing.
Proc. ACL 2012