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

Tree Automaton

Literature

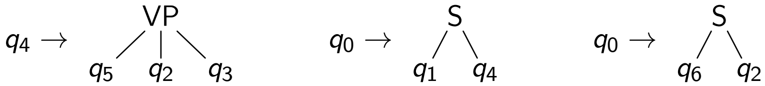
- THATCHER & WRIGHT (1968):
  Generalized Finite Automata Theory with an Application to a Decision Problem of Second-Order Logic.
  Math. Systems Theory 2(1): 57–81
- ROUNDS (1970):
  Mappings and Grammars on Trees.
  Mathematical Systems Theory 4(3): 257–287
- GÉCSEG (1977):
  Universal Algebras and Tree Automata.
  Proc. FCT: 98–112

tree automaton: (Q, Σ, I, R)

- finite set Q of states
- finite set Σ of terminals
- initial states $I \subseteq Q$
- finite set R of rules of the form $q \to \sigma(q_1, \dots, q_k)$

  $(\sigma \in \Sigma,\ k \geq 0,\ q, q_1, \dots, q_k \in Q)$

example rules:



Tree Automaton

derivation semantics: $\xi \Rightarrow \zeta$ for sentential forms $\xi, \zeta \in T_\Sigma(Q)$
if there exist leaf position w in ξ and rule $q \to r$ in R

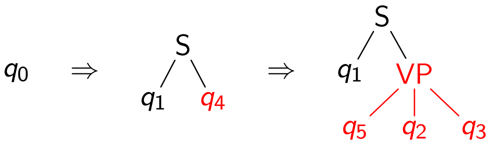
$$\xi = \xi[q]_w \qquad \zeta = \xi[r]_w$$

recognized tree language:

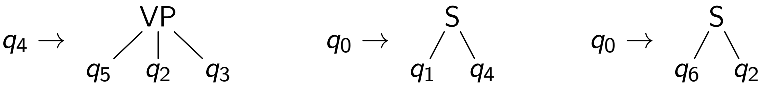
$$\{t \in T_\Sigma \mid \exists q_0 \in I \colon q_0 \Rightarrow^* t\}$$

Tree Automaton

example derivation:



example rules:



Tree Automaton

Standard references

GÉCSEG & STEINBY (1984):
Tree Automata.
Akadémiai Kiadó 2nd edition (2015): `arXiv:1509.06233`

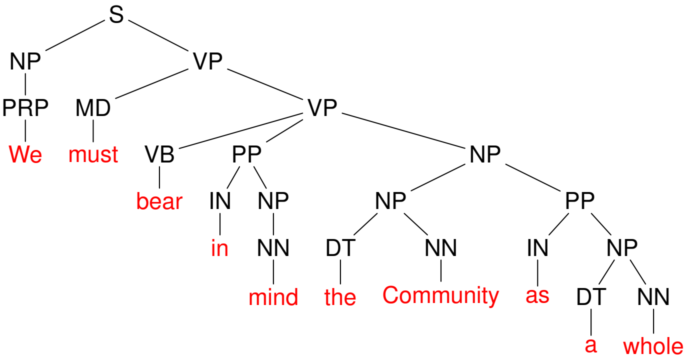
GÉCSEG & STEINBY (1997):
Tree Languages.
In *Handbook of Formal Languages* 3: 1–68, Springer

Parsing

Parsing

parsing: determining the syntactic structure of a sentence

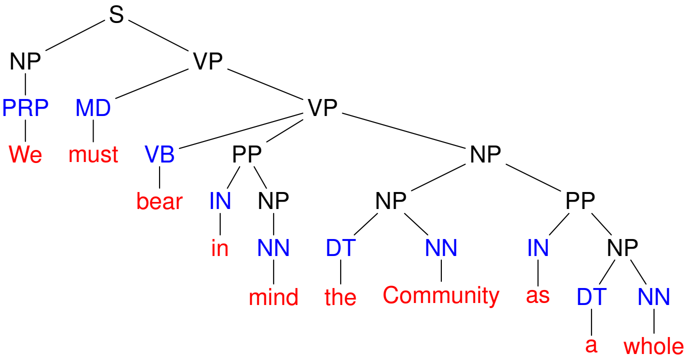
example: We must bear in mind the Community as a whole



Parsing

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

Parsing

short history:

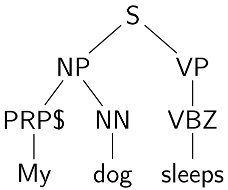
- before 1990
  - hand-crafted CFG (refined via POS tags)
  - corrections and selection by human annotators
- 1990s
  - PENN tree bank (1M words)
  - weighted CFG (statistic methods)
  - WALL STREET JOURNAL tree bank (30M words)
- since 2000
  - weighted tree automata

Parsing

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]

Parsing

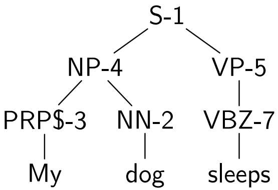


tags:

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

ADJA-Sup-Dat-Sg-Fem

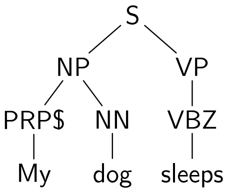
Parsing



tags:

- official tags often conservative
  - English: $\approx$ 50 tags
  - German: $\gg$ 200 tags — ADJA-Sup-Dat-Sg-Fem
- modern parsers use refined tags $\rightarrow$ subcategorization

Parsing



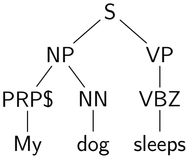
tags:

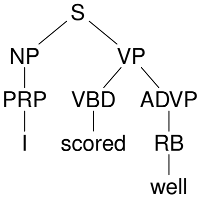
- official tags often conservative
  - English: $\approx$ 50 tags
  - German: $\gg$ 200 tags

ADJA-Sup-Dat-Sg-Fem

- modern parsers use refined tags $\rightarrow$ subcategorization
- but return parse over official tags $\rightarrow$ relabeling

Parsing

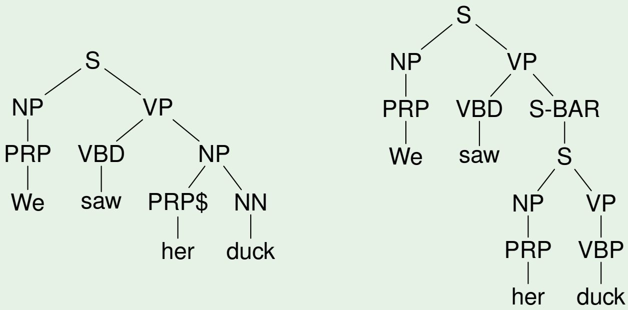




read off CFG productions:

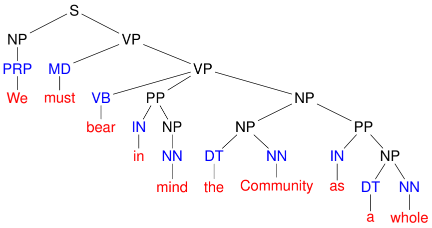
$$
\begin{array}{rclcrcl}
\text{S} & \longrightarrow & \text{NP VP} & \quad & \text{NP} & \longrightarrow & \text{PRP\$ NN} \\
\text{PRP\$} & \longrightarrow & \text{My} & & \text{NN} & \longrightarrow & \text{dog} \\
\text{VP} & \longrightarrow & \text{VBZ} & & \text{VBZ} & \longrightarrow & \text{sleeps} \\
\text{NP} & \longrightarrow & \text{PRP} & & \text{PRP} & \longrightarrow & \text{I} \\
\text{VP} & \longrightarrow & \text{VBD ADVP} & & \text{VBD} & \longrightarrow & \text{scored} \\
\text{ADVP} & \longrightarrow & \text{RB} & & \text{RB} & \longrightarrow & \text{well}
\end{array}
$$

Weights for disambiguation

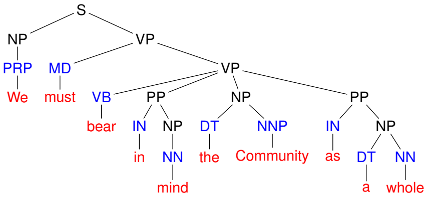


Parsing

BERKELEY parser [Reference]:



BLLIP parser:



Parsing

State-of-the-art models

- CFG with subcategorization (CFG_{sub})
- tree substitution grammars with subcategorization (TSG_{sub})
  [SHINDO et al., 2012]
- (both as expressive as weighted tree automata)
- other models

State-of-the-art models

- CFG with subcategorization (CFG_{sub})
- tree substitution grammars with subcategorization (TSG_{sub})
  [SHINDO et al., 2012]
- (both as expressive as weighted tree automata)
- other models

Experiment [SHINDO et al., 2012]

grammar model	F_1
wCFG	72.6
wTSG [COHN et al., 2010]	84.7
wCFG_{sub} [PETROV, 2010]	91.8
wTSG_{sub} [SHINDO et al., 2012]	92.4

Parsing

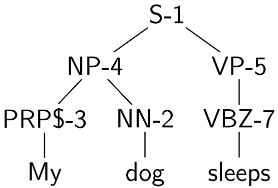
grammar with subcategorization: (grammar with relabeling)

- a grammar G generating $L(G) \subseteq T_\Sigma(W)$ (subcategorized trees)
- a (total) mapping $\rho\colon \Sigma \to \Delta$ (functional relabeling)

Parsing

grammar with subcategorization: (grammar with relabeling)

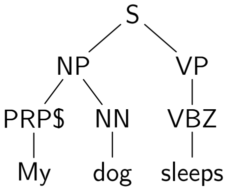
- a grammar G generating $L(G) \subseteq T_\Sigma(W)$ (subcategorized trees)
- a (total) mapping $\rho\colon \Sigma \to \Delta$ (functional relabeling)



Parsing

grammar with subcategorization: (grammar with relabeling)

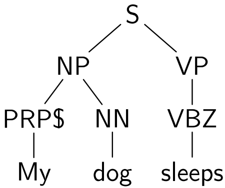
- a grammar G generating $L(G) \subseteq T_\Sigma(W)$ (subcategorized trees)
- a (total) mapping $\rho\colon \Sigma \to \Delta$ (functional relabeling)



Parsing

grammar with subcategorization: (grammar with relabeling)

- a grammar G generating $L(G) \subseteq T_\Sigma(W)$ (subcategorized trees)
- a (total) mapping $\rho \colon \Sigma \to \Delta$ (functional relabeling)



Semantics

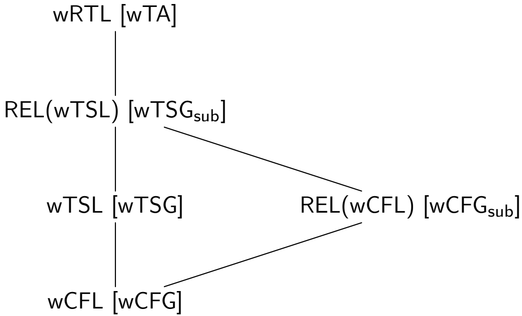
$$L(G, \rho) = \rho(L(G)) = \{\rho(t) \mid t \in L(G)\}$$

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

Parsing

Theorem

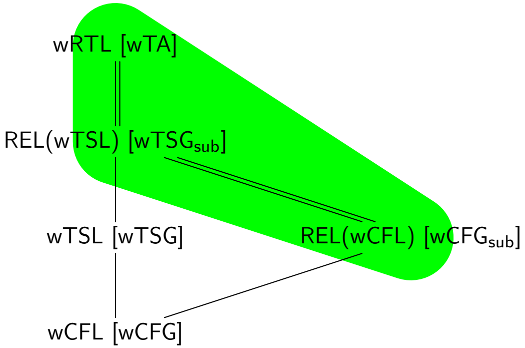
$\mathrm{REL}(\mathrm{wCFL}) = \mathrm{REL}(\mathrm{wTSL}) = \mathrm{wRTL}$



Parsing

Theorem

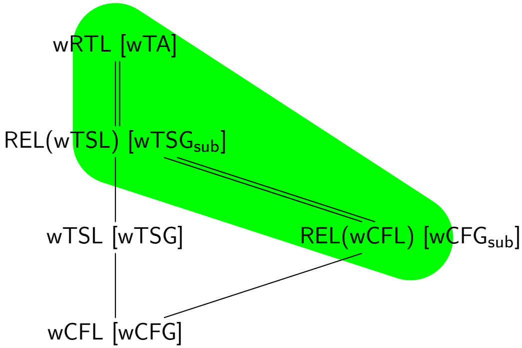
$\mathrm{REL}(\mathrm{wCFL}) = \mathrm{REL}(\mathrm{wTSL}) = \mathrm{wRTL}$



Parsing

Theorem

REL(wCFL) = REL(wTSL) = wRTL



hence: subcategorization ≈ finite-state

Parsing

Comparison:

1. rule of subcategorized grammar:

$$\text{S-1} \to \text{ADJP-2 S-1} \qquad \text{weight: } 0.003545$$

with relabeling $\rho(\text{S-1}) = \text{S}, \ldots$

Parsing

Comparison:

1. rule of subcategorized grammar:

$$\text{S-1} \to \text{ADJP-2 S-1} \qquad \text{weight: } 0.003545$$

   with relabeling $\rho(\text{S-1}) = \text{S}, \ldots$

2. corresponding rule of tree automaton

$$\text{S-1} \to \text{S}\big(\text{ADJP-2},\ \text{S-1}\big) \qquad \text{weight: } 0.003545$$

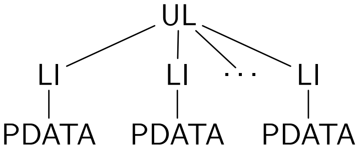
XML

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:

```
<UL>
<LI> first item </LI>
<LI> second item </LI>
...
<LI> last item </LI>
</UL>
```



XML

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

XML

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

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**