Foundations of Natural Language Processing
Lecture 13
Heads, Dependency parsing

Alex Lascarides
(slides from Alex Lascarides, Henry Thompson, Nathan Schneider and Sharon Goldwater)

6 March 2018
Today we will... 

- Provide metrics for evaluating a parser
- Return to the problem of PCFGs
- Suggest a fix
- This fix leads to an approach without constituent structure!

Dependency parsing
Evaluating parse accuracy

Compare gold standard tree (left) to parser output (right):

- Output constituent is counted correct if there is a gold constituent that spans the same sentence positions.
- Harsher measure: also require the constituent labels to match.
- Pre-terminals (lexical categories) don’t count as constituents.
Evaluating parse accuracy

Compare gold standard tree (left) to parser output (right):

- **Precision**: (# correct constituents)/(# in parser output) = 3/5
- **Recall**: (# correct constituents)/(# in gold standard) = 3/4
- **F-score**: balances precision/recall: 2pr/(p+r)
Parsing accuracies

F-scores for parsing on WSJ corpus:

- vanilla PCFG: $< 80\%$\(^1\)
- lexicalizing + cat-splitting: 89.5% (Charniak, 2000)
- Best current parsers get about 92%
- Numbers get better if we look at top 5 or top 10

However, results on other corpora and other languages are considerably lower. Definitely not a solved problem!

\(^{1}\)Charniak (1996) reports 81% but using gold POS tags as input.
• Probabilistic models of syntax can help disambiguation and speed in broad-coverage parsing.
  – by computing the probabilities of each tree or sub-tree as the product of the rules in it, and choosing the best option(s).

• Treebanks provide training data for estimating rule probabilities.

• However, to do well, we need to be clever:
  – Standard categories in the treebank don’t capture some important facts about language.
  – By creating more detailed categories, we can encode more information within the PCFG framework.
Recall Problem with Vanilla PCFGs
No lexical dependencies

Replacing one word with another with the same POS will never result in a different parsing decision, even though it should!

- kids saw birds with fish vs. kids saw birds with binoculars
- She stood by the door covered in tears vs. She stood by the door covered in ivy
- stray cats and dogs vs. Siamese cats and dogs
A way to fix PCFGs: lexicalization

Create new categories, this time by adding the **lexical head** of the phrase (note: N level under NPs not shown for brevity):

```
S-saw
   NP-kids        VP-saw
      kids         
        VP-saw    PP-binoculars
           V-saw  NP-birds   P-with  NP-binoculars
               saw   birds  with   binoculars
```

- Now consider:

  \[
  \text{VP-saw} \rightarrow \text{VP-saw PP-fish vs. VP-saw} \rightarrow \text{VP-saw PP-binoculars}
  \]
Practical issues

• All this category-splitting makes the grammar much more specific (good!)

• But leads to huge grammar blowup and very sparse data (bad!)

• Lots of effort on how to balance these two issues.
  – Complex smoothing schemes (similar to N-gram interpolation/backoff).
  – More recently, increasing emphasis on automatically learned subcategories.

• But do we really need phrase structure in the first place? Not always!

• Today: Syntax (and parsing) without constituent structure.
Outline

1. Dependencies: what/why

2. Transforming constituency → dependency parse

3. Direct dependency parsing
   • Transition-based (shift-reduce)
   • Graph-based
Lexicalized Constituency Parse

```
S-saw
  NP-kids
    kids
  VP-saw
    V-saw
      saw
    NP-birds
      NP-birds
        birds
      PP-fish
        P-with
          with
        NP-fish
          fish
```
. . . remove the phrasal categories. . .
... remove the (duplicated) terminals. ...
and collapse chains of duplicates.
... and collapse chains of duplicates. ...

```
saw
  kids
    saw
      birds
        fish
        with
```
... and collapse chains of duplicates. ...
... and collapse chains of duplicates. ...
... and collapse chains of duplicates. ...

```
saw

kids  saw

saw  birds
    |
    fish
    |
    with
```
... and collapse chains of duplicates. ...

saw
  
  kids  birds
      |
      fish
      |
      with
Linguists have long observed that the meanings of words within a sentence depend on one another, mostly in asymmetric, binary relations.

- Though some constructions don’t cleanly fit this pattern: e.g., coordination, relative clauses.
Equivalently, but showing word order (head $\rightarrow$ modifier):

Because it is a tree, every word has exactly one parent.
Content vs. Functional Heads

Some treebanks prefer **content heads**:

```
Little kids were always watching birds with fish
```

Others prefer **functional heads**:

```
Little kids were always watching birds with fish
```
It is often useful to distinguish different kinds of head → modifier relations, by labeling edges:

Important relations for English include subject, direct object, determiner, adjective modifier, adverbial modifier, etc. (Different treebanks use somewhat different label sets.)

- How would you identify the subject in a constituency parse?
Dependency Paths

For information extraction tasks involving real-world relationships between entities, chains of dependencies can provide good features:

(Example from Brendan O’Connor)
Projectivity

- A sentence’s dependency parse is said to be **projective** if every subtree (node and all its descendants) occupies a *contiguous span* of the sentence.

- The dependency parse can be drawn on top of the sentence without any crossing edges.

```
A hearing on the issue is scheduled today
```
Other sentences are nonprojective:

Nonprojectivity is rare in English, but quite common in many languages.
Outline

1. Dependencies: what/why

2. **Transforming constituency → dependency parse**

3. Direct dependency parsing
   - Transition-based (shift-reduce)
   - Graph-based
Constituency Tree → Dependency Tree

We saw how the **lexical head** of the phrase can be used to collapse down to a dependency tree:

```
S-saw
  +---+---+
  |   |   |
  NP-kids VP-saw
        +---+---+
        |   |   |
        kids VP-saw
            +---+---+
            |   |   |
            V-saw PP-binoculars
                +---+---+
                |   |   |
                NP-birds P-with NP-binoculars
                        +---+---+
                        |   |   |
                        saw with binoculars
```

- But how can we find each phrase's head in the first place?
**Head Rules**

The standard solution is to use **head rules**: for every non-unary (P)CFG production, designate one RHS nonterminal as containing the head. $S \rightarrow \text{NP } \text{VP}$, $\text{VP} \rightarrow \text{VP } \text{PP}$, $\text{PP} \rightarrow \text{P } \text{NP}$ (content head), etc.

- Heuristics to scale this to large grammars: e.g., within an NP, last immediate N child is the head.
Then, propagate heads up the tree:

$$\text{S}$$

- $$\text{NP-kids}$$
  - kids

$$\text{VP}$$

- $$\text{V-saw}$$
  - saw
- $$\text{NP-birds}$$
  - birds

$$\text{PP}$$

- $$\text{P-with}$$
  - with
- $$\text{NP-binoculars}$$
  - binoculars
Head Rules

Then, propagate heads up the tree:

```
S
  NP-kids
    kids
  VP-saw
    V-saw
      saw
    NP-birds
      birds
  PP
    P-with
      with
    NP-binoculars
      binoculars
```

Alex Lascarides

FNLP Lecture 13
Head Rules

Then, propagate heads up the tree:

\[
\begin{align*}
S & \quad VP \\
\text{NP-kids} & \quad \text{VP-saw} \\
& \quad \text{V-saw} \\
& \quad \text{saw} \\
\text{kids} & \quad \text{NP-birds} \\
& \quad \text{PP-binoculars} \\
& \quad \text{P-with} \\
& \quad \text{with} \\
& \quad \text{NP-binoculars} \\
& \quad \text{binoculars}
\end{align*}
\]
Head Rules

Then, propagate heads up the tree:

```
S
  ___________
 NP-kids      VP-saw
    |         |
   kids       PP-binoculars
   ___________
 VP-saw      VP-saw
   _____     _____
 V-saw      V-saw
     |       |     |
    saw     saw     |
    |       |     |
   NP-birds NP-birds |
     |     |
    with with       |
    |     |
   NP-binoculars  binoculars
```
Head Rules

Then, propagate heads up the tree:

```
S-saw
  NP-kids
    kids
  VP-saw
    VP-saw
      V-saw
        saw
      NP-birds
        birds
    PP-binoculars
      P-with
        with
      NP-binoculars
        binoculars
```
Outline

1. Dependencies: what/why

2. Transforming constituency $\rightarrow$ dependency parse

3. Direct dependency parsing
   - Transition-based (shift-reduce)
   - Graph-based
Dependency Parsing

Some of the algorithms you have seen for PCFGs can be adapted to dependency parsing.

- **CKY** can be adapted, though efficiency is a concern: obvious approach is $O(Gn^5)$; Eisner algorithm brings it down to $O(Gn^3)$

- **Shift-reduce**: more efficient, doesn’t even require a grammar!
Transitation-based Parsing: Shift Reduce Parser

3 possible actions:

**LeftArc**: Assign head-dependent relation between $s_1$ and $s_2$; pop $s_2$

**RightArc**: Assign head-dependent relation between $s_2$ and $s_1$; pop $s_1$

**Shift**: Put $w_1$ on top of the stack.
### Example

<table>
<thead>
<tr>
<th>Step</th>
<th>Stack</th>
<th>Word List</th>
<th>Action</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>[root]</td>
<td>[Kim,saw,Sandy]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>[root,Kim]</td>
<td>[saw,Sandy]</td>
<td>Shift</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>[root,Kim,saw]</td>
<td>[Sandy]</td>
<td>Shift</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>[root,saw]</td>
<td>[Sandy]</td>
<td>LeftArc</td>
<td>nsubj(saw,Kim)</td>
</tr>
<tr>
<td>4</td>
<td>[root,saw,Sandy]</td>
<td></td>
<td>Shift</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>[root,saw]</td>
<td></td>
<td>RightArc</td>
<td>dobj(saw,Sandy)</td>
</tr>
<tr>
<td>6</td>
<td>[root]</td>
<td></td>
<td>RightArc</td>
<td>root→book</td>
</tr>
</tbody>
</table>

**Diagram:**

```
ROOT

<table>
<thead>
<tr>
<th>NSUBJ</th>
<th>DOBJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim</td>
<td>saw</td>
</tr>
<tr>
<td></td>
<td>Sandy</td>
</tr>
</tbody>
</table>
```
Transition-based Parsing

- Latent structure is just edges between words. Train a classifier to predict next action (SHIFT, REDUCE, ATTACH-LEFT, or ATTACH-RIGHT), and proceed left-to-right through the sentence. $O(n)$ time complexity!

- Only finds projective trees (without special extensions)

- Pioneering system: Nivre’s MALTPARSER

- See [http://spark-public.s3.amazonaws.com/nlp/slides/Parsing-Dependency.pdf](http://spark-public.s3.amazonaws.com/nlp/slides/Parsing-Dependency.pdf) (Jurafsky & Manning Coursera slides) for details and examples
Graph-based Parsing

• Global algorithm: From the fully connected directed graph of all possible edges, choose the best ones that form a tree.

• Edge-factored models: Classifier assigns a nonnegative score to each possible edge; maximum spanning tree algorithm finds the spanning tree with highest total score in $O(n^2)$ time.
  
  – Edge-factored assumption can be relaxed (higher-order models score larger units; more expensive).
  – Unlabeled parse $\rightarrow$ edge-labeling classifier (pipeline).

• Pioneering work: McDonald’s MSTParser

• Can be formulated as constraint-satisfaction with integer linear programming (Martins’s TurboParser)
Graph-based vs. Transition-based vs. Conversion-based

- TB: Features in scoring function can look at any part of the stack; no optimality guarantees for search; linear-time; (classically) projective only

- GB: Features in scoring function limited by factorization; optimal search within that model; quadratic-time; no projectivity constraint

- CB: In terms of accuracy, sometimes best to first constituency-parse, then convert to dependencies (e.g., Stanford Parser). Slower than direct methods.
Choosing a Parser: Criteria

- Target representation: constituency or dependency?
- Efficiency? In practice, both runtime and memory use.
- Incrementality: parse the whole sentence at once, or obtain partial left-to-right analyses/expectations?
- Retrainable system?
Choosing a Parser: Performance

SOTA for English constituency parsing (WSJ §23): 91%–92% $F_1$

<table>
<thead>
<tr>
<th>Parser</th>
<th>LR</th>
<th>LP</th>
<th>$F_1$</th>
<th>#Toks/s.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charniak (2000)</td>
<td>89.5</td>
<td>89.9</td>
<td>89.5</td>
<td>–</td>
</tr>
<tr>
<td>Klein and Manning (2003)</td>
<td>85.3</td>
<td>86.5</td>
<td>85.9</td>
<td>143</td>
</tr>
<tr>
<td>Petrov and Klein (2007)</td>
<td>90.0</td>
<td>90.3</td>
<td>90.1</td>
<td>169</td>
</tr>
<tr>
<td>Carreras et al. (2008)</td>
<td>90.7</td>
<td>91.4</td>
<td>91.1</td>
<td>–</td>
</tr>
<tr>
<td>Zhu et al. (2013)</td>
<td>90.3</td>
<td>90.6</td>
<td>90.4</td>
<td>1,290</td>
</tr>
<tr>
<td>Stanford Shift-Reduce (2014)</td>
<td>89.1</td>
<td>89.1</td>
<td>89.1</td>
<td>655</td>
</tr>
<tr>
<td>Hall et al. (2014)</td>
<td>88.4</td>
<td>88.8</td>
<td>88.6</td>
<td>12</td>
</tr>
<tr>
<td><strong>This work</strong></td>
<td>89.9</td>
<td>90.4</td>
<td>90.2</td>
<td>957</td>
</tr>
<tr>
<td>Charniak and Johnson (2005)*</td>
<td>91.2</td>
<td>91.8</td>
<td>91.5</td>
<td>84</td>
</tr>
<tr>
<td>Socher et al. (2013)*</td>
<td>89.1</td>
<td>89.7</td>
<td>89.4</td>
<td>70</td>
</tr>
<tr>
<td>Zhu et al. (2013)*</td>
<td>91.1</td>
<td>91.5</td>
<td>91.3</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 3: Results on the English PTB §23. All systems reporting runtimes were run on the same machine. Marked as * are reranking and semi-supervised c-parsers.

(Fernández-González and Martins, 2015)
Choosing a Parser: Performance

Constituency parsing in other languages

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley</td>
<td>70.50</td>
<td>80.38</td>
<td>78.30</td>
<td>86.96</td>
<td>81.62</td>
<td>71.42</td>
<td>79.23</td>
<td>79.19</td>
<td>78.45</td>
</tr>
<tr>
<td>Berkeley Tagged</td>
<td>74.74</td>
<td>79.76</td>
<td>78.28</td>
<td>85.42</td>
<td>85.22</td>
<td>78.56</td>
<td>86.75</td>
<td>80.64</td>
<td>81.17</td>
</tr>
<tr>
<td>Hall et al. (2014)</td>
<td>83.39</td>
<td>79.70</td>
<td>78.43</td>
<td>87.18</td>
<td>88.25</td>
<td>80.18</td>
<td>90.66</td>
<td>82.00</td>
<td>83.72</td>
</tr>
<tr>
<td>Crabbé and Seddah (2014)</td>
<td>85.35</td>
<td>79.68</td>
<td>77.15</td>
<td>86.19</td>
<td>87.51</td>
<td>79.35</td>
<td>91.60</td>
<td>82.72</td>
<td>83.69</td>
</tr>
<tr>
<td>This work</td>
<td>85.90</td>
<td>78.75</td>
<td>78.66</td>
<td>88.97</td>
<td>88.16</td>
<td>79.28</td>
<td>91.20</td>
<td>82.80</td>
<td>84.22</td>
</tr>
<tr>
<td>Björkelund et al. (2014)</td>
<td>88.24</td>
<td>82.53</td>
<td>81.66</td>
<td>89.80</td>
<td>91.72</td>
<td>83.81</td>
<td>90.50</td>
<td>85.50</td>
<td>86.72</td>
</tr>
</tbody>
</table>

(Fernández-González and Martins, 2015)
Choosing a Parser: Performance

SOTA for English dependency parsing (WSJ §23): 93%–94% UAS, 91%–92% LAS

<table>
<thead>
<tr>
<th>System</th>
<th>UAS</th>
<th>LAS</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline greedy parser</td>
<td>91.47</td>
<td>90.43</td>
<td>0.001</td>
</tr>
<tr>
<td>Huang and Sagae (2010)</td>
<td>92.10</td>
<td>90.43</td>
<td>0.04</td>
</tr>
<tr>
<td>Zhang and Nivre (2011)</td>
<td>92.90</td>
<td>91.80</td>
<td>0.03</td>
</tr>
<tr>
<td>Choi and McCallum (2013)</td>
<td>92.96</td>
<td>91.93</td>
<td>0.009</td>
</tr>
<tr>
<td>Ma et al. (2014)</td>
<td>93.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bohnet and Nivre (2012)†‡</td>
<td>93.67</td>
<td>92.68</td>
<td>0.4</td>
</tr>
<tr>
<td>Suzuki et al. (2009)†</td>
<td>93.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koo et al. (2008)†</td>
<td>93.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen et al. (2014)†</td>
<td>93.77</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| beam size                     |       |       |       |
| training decoding             |       |       |       |
| 100  100                       | 93.28 | 92.35 | 0.07  |
| 100  64                        | 93.20 | 92.27 | 0.04  |
| 100  16                        | 92.40 | 91.95 | 0.01  |

Table 5: Results on WSJ. Speed: sentences per second. †: semi-supervised learning. ‡: joint POS-tagging and dependency parsing models.

(Zhou et al., 2015)
Summary

• While constituency parses give hierarchically nested phrases, dependency parses represent syntax with trees whose edges connect words in the sentence. (No abstract phrase categories like NP.) Edges often labeled with relations like subject.

• Head rules govern how a lexicalized constituency grammar can be extracted from a treebank, and how a constituency parse can be converted to a dependency parse.

• For English, it is often fastest and most convenient to parse directly to dependencies. Two main paradigms, graph-based and transition-based, with different kinds of models and search algorithms.

• Google “online dependency parser”.
  Try out the Stanford parser and SEMAFOR!