Log-Linear Models with Structured Outputs

Natural Language Processing
CS 6120—Spring 2014
Northeastern University

David Smith
(some slides from Andrew McCallum)
Overview

• Sequence labeling task (cf. POS tagging)
• Independent classifiers
• HMMs
• (Conditional) Maximum Entropy Markov Models
• Conditional Random Fields
• Beyond Sequence Labeling
Sequence Labeling

• Inputs: $x = (x_1, \ldots, x_n)$
• Labels: $y = (y_1, \ldots, y_n)$
• Typical goal: Given $x$, predict $y$

• Example sequence labeling tasks
  – Part-of-speech tagging
  – Named-entity-recognition (NER)
    • Label people, places, organizations
Red Sox and Their Fans Let Loose

Fans of the slugger David Ortiz in Boston’s Copley Square.

By PETE THAMEL
Published: October 31, 2007

BOSTON Oct. 30 — Jonathan Papelbon turned Boston’s World Series victory parade into a full-scale dance party Tuesday as the Red Sox put an exclamation point on the 2007 season.
First Solution: Maximum Entropy Classifier

• Conditional model $p(y|x)$.
  – Do not waste effort modeling $p(x)$, since $x$ is given at test time anyway.
  – Allows more complicated input features, since we do not need to model dependencies between them.

• Feature functions $f(x,y)$:
  – $f_1(x,y) = \{ \text{word is Boston} \land y=\text{Location} \}$
  – $f_2(x,y) = \{ \text{first letter capitalized} \land y=\text{Name} \}$
  – $f_3(x,y) = \{ x \text{ is an HTML link} \land y=\text{Location} \}$
First Solution: MaxEnt Classifier

• How should we choose a classifier?

• Principle of maximum entropy
  – We want a classifier that:
    • Matches feature constraints from training data.
    • Predictions maximize entropy.

• There is a unique, exponential family distribution that meets these criteria.
First Solution: MaxEnt Classifier

• Problem with using a maximum entropy classifier for sequence labeling:
• It makes decisions at each position independently!
Second Solution: HMM

\[ P(y, x) = \prod_{t} P(y_t \mid y_{t-1})P(x \mid y_t) \]

- Defines a generative process.
- Can be viewed as a weighted finite state machine.
Second Solution: HMM

- How can we represent multiple features in an HMM?
  - Treat them as conditionally independent given the class label?
    - The example features we talked about are not independent.
  - Try to model a more complex generative process of the input features?
    - We may lose tractability (i.e. lose a dynamic programming for exact inference).
Second Solution: HMM

• Let’s use a conditional model instead.
Third Solution: MEMM

- Use a series of maximum entropy classifiers that know the previous label.
- Define a Viterbi algorithm for inference.

\[ P(y \mid x) = \prod_t P_{y_{t-1}}(y_t \mid x) \]
Third Solution: MEMM

• Combines the advantages of maximum entropy and HMM!
• But there is a problem…
Problem with MEMMs: Label Bias

- In some state space configurations, MEMMs essentially completely ignore the inputs.

- This is not a problem for HMMs, because the input sequence is generated by the model.
Fourth Solution: Conditional Random Field

- Conditionally-trained, undirected graphical model.
- For a standard linear-chain structure:

\[ P(y \mid x) = \prod_t \Psi_k(y_t, y_{t-1}, x) \]

\[ \Psi_k(y_t, y_{t-1}, x) = \exp \left( \sum_k \lambda_k f(y_t, y_{t-1}, x) \right) \]
Fourth Solution: CRF

• Have the advantages of MEMMs, but avoid the label bias problem.

• CRFs are globally normalized, whereas MEMMs are locally normalized.

• Widely used and applied. CRFs give state-the-art results in many domains.
Fourth Solution: CRF

- Have the advantages of MEMMs, but avoid the label bias problem.

- CRFs are globally normalized, whereas MEMMs are locally normalized.

- Widely used and applied, state-of-the-art results in many applications.

Remember, $Z$ is the normalization constant. How do we compute it?
CRF Applications

- Part-of-speech tagging
- Named entity recognition
- Document layout (e.g. table) classification
- Gene prediction
- Chinese word segmentation
- Morphological disambiguation
- Citation parsing
- Etc., etc.
NER as Sequence Tagging

The Phoenicians came from the Red Sea
NER as Sequence Tagging

The Phoenicians came from the Red Sea
NER as Sequence Tagging

The Phoenicians came from the Red Sea
NER as Sequence Tagging

The Phoenicians came from the Red Sea

- Capitalized word
- Ends in “s”

The Phoenicians  came  from  the  Red  Sea
The Phoenicians came from the Red Sea
The Phoenicians came from the Red Sea
The Phoenicians came from the Red Sea
NER as Sequence Tagging

The Phoenicians came from the Red Sea
NER as Sequence Tagging

The Phoenicians came from the Red Sea

Not capitalized
NER as Sequence Tagging

The Phoenicians came from the Red Sea.
NER as Sequence Tagging

The Phoenicians came from the Red Sea

- B-E to right
- Not capitalized
- Tagged as “VB”
NER as Sequence Tagging

The Phoenicians came from the Red Sea
NER as Sequence Tagging

The Phoenicians came from the Red Sea
The Phoenicians came from the Red Sea
NER as Sequence Tagging

The Phoenicians came from the Red Sea

Word “sea” preceded by “the ADJ”

Word “sea”

Hard constraint: I-L must follow B-L or I-L
Overview

• What computations do we need?
• Smoothing log-linear models
• MEMMs vs. CRFs again
  • Action-based parsing and dependency parsing
Recipe for Conditional Training of \( p(y \mid x) \)

1. Gather constraints/features from training data
   \[
   \alpha_{iy} = \tilde{E}[f_{iy}] = \sum_{x_j, y_j \in D} f_{iy}(x_j, y_j)
   \]

2. Initialize all parameters to zero

3. Classify training data with current parameters; calculate expectations
   \[
   E_\Theta[f_{iy}] = \sum_{x_j \in D} \sum_{y' \in D} p_\Theta(y' \mid x_j) f_{iy}(x_j, y')
   \]

4. Gradient is
   \[
   \tilde{E}[f_{iy}] - E_\Theta[f_{iy}]
   \]

5. Take a step in the direction of the gradient

6. Repeat from 3 until convergence
Recipe for Conditional Training of $p(y \mid x)$

1. Gather constraints/features from training data
   $$\alpha_{iy} = \tilde{E}[f_{iy}] = \sum_{x_j, y_j \in D} f_{iy}(x_j, y_j)$$

2. Initialize all parameters to zero

3. Classify training data with current parameters; calculate expectations
   $$E_\Theta[f_{iy}] = \sum_{x_j \in D} \sum_{y'} p_\Theta(y' \mid x_j) f_{iy}(x_j, y')$$

4. Gradient is
   $$\tilde{E}[f_{iy}] - E_\Theta[f_{iy}]$$

5. Take a step in the direction of the gradient

6. Repeat from 3 until convergence

Where have we seen expected counts before?
Recipe for Conditional Training of \( p(y | x) \)

1. Gather constraints/features from training data
   \[ \alpha_{i,y} = \bar{E}[f_{i,y}] = \sum_{x_j, y_j \in D} f_{i,y}(x_j, y_j) \]

2. Initialize all parameters to zero

3. Classify training data with current parameters; calculate expectations
   \[ E_\Theta[f_{i,y}] = \sum_{x_j \in D} \sum_{y' \in y'} p_\Theta(y' | x_j) f_{i,y}(x_j, y') \]

4. Gradient is
   \[ \bar{E}[f_{i,y}] - E_\Theta[f_{i,y}] \]

5. Take a step in the direction of the gradient

6. Repeat from 3 until convergence

Where have we seen expected counts before?
Gradient-Based Training

- $\lambda := \lambda + \text{rate} \times \text{Gradient}(F)$
- After all training examples? (batch)
- After every example? (on-line)
- Use second derivative for faster learning?
- A big field: numerical optimization
Overfitting

• If we have too many features, we can choose weights to model the training data perfectly.

• If we have a feature that only appears in spam training, not ham training, it will get weight $\infty$ to maximize $p(\text{spam} \mid \text{feature})$ at 1.

• These behaviors
  • Overfit the training data
  • Will probably do poorly on test data
Solutions to Overfitting

• Throw out rare features.
  • Require every feature to occur > 4 times, and > 0 times with
    ling, and > 0 times with spam.

• Only keep, e.g., 1000 features.
  • Add one at a time, always greedily picking the one that most
    improves performance on held-out data.

• Smooth the observed feature counts.

• Smooth the weights by using a prior.
  • \( \max p(\lambda | \text{data}) = \max p(\lambda, \text{data}) = p(\lambda) p(\text{data} | \lambda) \)
  • decree \( p(\lambda) \) to be high when most weights close to 0
Smoothing with Priors

• What if we had a prior expectation that parameter values wouldn’t be very large?

• We could then balance evidence suggesting large (or infinite) parameters against our prior expectation.

• The evidence would never totally defeat the prior, and parameters would be smoothed (and kept finite)

• We can do this explicitly by changing the optimization objective to maximum posterior likelihood:

\[ \log P(y, \lambda | x) = \log P(\lambda) + \log P(y | x, \lambda) \]

Posterior  Prior  Likelihood
Smoothing: Priors

- Gaussian, or quadratic, priors:
  - Intuition: parameters shouldn’t be large.
  - Formalization: prior expectation that each parameter will be distributed according to a gaussian with mean $\mu$ and variance $\sigma^2$.

  $$P(\lambda_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(\lambda_i - \mu_i)^2}{2\sigma_i^2}\right)$$

- Penalizes parameters for drifting to far from their mean prior value (usually $\mu=0$).
- $2\sigma^2=1$ works surprisingly well.
Parsing as Structured Prediction
## Shift-reduce parsing

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input remaining</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>()</td>
<td>Book that flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Book)</td>
<td>that flight</td>
<td>reduce, Verb → book, (Choice #1 of 2)</td>
</tr>
<tr>
<td>(Verb)</td>
<td>that flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Verb that)</td>
<td>flight</td>
<td>reduce, Det → that</td>
</tr>
<tr>
<td>(Verb Det)</td>
<td>flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Verb Det flight)</td>
<td></td>
<td>reduce, Noun → flight</td>
</tr>
<tr>
<td>(Verb Det Noun)</td>
<td></td>
<td>reduce, NOM → Noun</td>
</tr>
<tr>
<td>(Verb Det NOM)</td>
<td></td>
<td>reduce, NP → Det NOM</td>
</tr>
<tr>
<td>(Verb NP)</td>
<td></td>
<td>reduce, VP → Verb NP</td>
</tr>
<tr>
<td>(Verb)</td>
<td></td>
<td>reduce, S → V</td>
</tr>
<tr>
<td>(S)</td>
<td></td>
<td>SUCCESS!</td>
</tr>
</tbody>
</table>

Ambiguity may lead to the need for backtracking.
### Shift-reduce parsing

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input remaining</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>()</td>
<td>Book that flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Book)</td>
<td>that flight</td>
<td>reduce, Verb → book, (Choice #1 of 2)</td>
</tr>
<tr>
<td>(Verb)</td>
<td>that flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Verb that)</td>
<td>flight</td>
<td>reduce, Det → that</td>
</tr>
<tr>
<td>(Verb Det)</td>
<td>flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Verb Det flight)</td>
<td>flight</td>
<td>reduce, Noun → flight</td>
</tr>
<tr>
<td>(Verb Det Noun)</td>
<td></td>
<td>reduce, NOM → Noun</td>
</tr>
<tr>
<td>(Verb Det NOM)</td>
<td></td>
<td>reduce, NP → Det NOM</td>
</tr>
<tr>
<td>(Verb NP)</td>
<td></td>
<td>reduce, VP → Verb NP</td>
</tr>
<tr>
<td>(Verb)</td>
<td></td>
<td>reduce, S → V</td>
</tr>
<tr>
<td>(S)</td>
<td></td>
<td>SUCCESS!</td>
</tr>
</tbody>
</table>

Ambiguity may lead to the need for backtracking.
### Shift-reduce parsing

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input remaining</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>()</td>
<td>Book that flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Book)</td>
<td>that flight</td>
<td>reduce, Verb → book, (Choice #1 of 2)</td>
</tr>
<tr>
<td>(Verb)</td>
<td>that flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Verb that)</td>
<td>flight</td>
<td>reduce, Det → that</td>
</tr>
<tr>
<td>(Verb Det)</td>
<td>flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Verb Det flight)</td>
<td></td>
<td>reduce, Noun → flight</td>
</tr>
<tr>
<td>(Verb Det Noun)</td>
<td></td>
<td>reduce, NOM → Noun</td>
</tr>
<tr>
<td>(Verb Det NOM)</td>
<td></td>
<td>reduce, NP → Det NOM</td>
</tr>
<tr>
<td>(Verb NP)</td>
<td></td>
<td>reduce, VP → Verb NP</td>
</tr>
<tr>
<td>(Verb)</td>
<td></td>
<td>reduce, S → V</td>
</tr>
<tr>
<td>(S)</td>
<td></td>
<td>SUCCESS!</td>
</tr>
</tbody>
</table>

Ambiguity may lead to the need for backtracking.

Train log-linear model of $p(\text{action} \mid \text{context})$
**Shift-reduce parsing**

<table>
<thead>
<tr>
<th>Stack</th>
<th>Input remaining</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>()</td>
<td>Book that flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Book)</td>
<td>that flight</td>
<td>reduce, Verb → book, (Choice #1 of 2)</td>
</tr>
<tr>
<td>(Verb)</td>
<td>that flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Verb that)</td>
<td>flight</td>
<td>reduce, Det → that</td>
</tr>
<tr>
<td>(Verb Det)</td>
<td>flight</td>
<td>shift</td>
</tr>
<tr>
<td>(Verb Det flight)</td>
<td></td>
<td>reduce, Noun → flight</td>
</tr>
<tr>
<td>(Verb Det Noun)</td>
<td></td>
<td>reduce, NOM → Noun</td>
</tr>
<tr>
<td>(Verb Det NOM)</td>
<td></td>
<td>reduce, NP → Det NOM</td>
</tr>
<tr>
<td>(Verb NP)</td>
<td></td>
<td>reduce, VP → Verb NP</td>
</tr>
<tr>
<td>(Verb)</td>
<td></td>
<td>reduce, S → V</td>
</tr>
<tr>
<td>(S)</td>
<td></td>
<td>SUCCESS!</td>
</tr>
</tbody>
</table>

Ambiguity may lead to the need for backtracking.

Train log-linear model of $p(\text{action} | \text{context})$
Structured Log-Linear Models
Structured Log-Linear Models

- Linear model for scoring structures

\[ \text{score}(out, in) = \theta \cdot \text{features}(out, in) \]
Structured Log-Linear Models

• Linear model for scoring structures
• Get a probability distribution by normalizing

\[
\text{score}(out, in) = \theta \cdot \text{features}(out, in)
\]

\[
p(out \mid in) = \frac{1}{Z} e^{\text{score}(out,in)} \quad Z = \sum_{out' \in GEN(in)} e^{\text{score}(out',in)}
\]
Structured Log-Linear Models

- Linear model for scoring structures
- Get a probability distribution by normalizing
  - Viz. logistic regression, Markov random fields, undirected graphical models

\[
\begin{align*}
\text{score}(\text{out}, \text{in}) &= \theta \cdot \text{features}(\text{out}, \text{in}) \\
p(\text{out} \mid \text{in}) &= \frac{1}{Z} e^{\text{score}(\text{out}, \text{in})} \\
Z &= \sum_{\text{out}' \in \text{GEN}(\text{in})} e^{\text{score}(\text{out}', \text{in})}
\end{align*}
\]
Structured Log-Linear Models

- Linear model for scoring structures
- Get a probability distribution by normalizing
  - Viz. logistic regression, Markov random fields, undirected graphical models

$$p(out \mid in) = \frac{1}{Z} e^{score(out,in)}$$

$$Z = \sum_{out' \in GEN(in)} e^{score(out',in)}$$

**Usually the bottleneck in NLP**

$$score(out, in) = \theta \cdot \text{features}(out, in)$$
Structured Log-Linear Models

• Linear model for scoring structures
• Get a probability distribution by normalizing
  ✤ Viz. logistic regression, Markov random fields, undirected graphical models
• Inference: sampling, variational methods, dynamic programming, local search, ...

\[
p(out \mid in) = \frac{1}{Z} e^{\text{score}(out,in)} \quad Z = \sum_{out' \in GEN(in)} e^{\text{score}(out',in)}
\]
Structured Log-Linear Models

- Linear model for scoring structures
- Get a probability distribution by normalizing
  - Viz. logistic regression, Markov random fields, undirected graphical models
- Inference: sampling, variational methods, dynamic programming, local search, ...
- Training: maximum likelihood, minimum risk, etc.

$$p(out \mid in) = \frac{1}{Z} e^{score(out, in)} \quad Z = \sum_{out' \in GEN(in)} e^{score(out', in)}$$

Usually the bottleneck in NLP

$$score(out, in) = \theta \cdot features(out, in)$$
Structured Log-Linear Models

With latent variables

• Several layers of linguistic structure
• Unknown correspondences
• Naturally handled by probabilistic framework
• Several inference setups, for example:

\[
p(out_1 \mid in) = \sum_{out_2, alignment} p(out_1, out_2, alignment \mid in)
\]
Structured Log-Linear Models

With latent variables

- Several layers of linguistic structure
- Unknown correspondences
- Naturally handled by probabilistic framework
- Several inference setups, for example:

\[ p(out_1 \mid in) = \sum_{out_2, alignment} p(out_1, out_2, alignment \mid in) \]
Edge-Factored Parsers

- No global features of a parse (McDonald et al. 2005)
- Each feature is attached to some edge
- MST or CKY-like DP for fast $O(n^2)$ or $O(n^3)$ parsing

“IT was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Is this a good edge?

Byl jasný studený dubnový den a hodiny odbíjely třináctou

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Is this a good edge?

yes, lots of positive features ...

Byl jasný studený dubnový den a hodiny odbíjely třináctou

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Is this a good edge?

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Is this a good edge?

jasný ⇐ den
(“bright day”)

Byl jasný studený dubnový den a hodiny odbíjely třináctou

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Is this a good edge?

jasný ← den
(“bright day”)

“...a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Is this a good edge?

jasný ← den
(“bright day”)

jasný ← N
(“bright NOUN”)

Byl jasný studený dubnový den a hodiny odbíjely třináctou

V A A A A N J N V C

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Is this a good edge?

jasný \leftarrow \text{den}  
(“bright day”)  

jasný \leftarrow \text{N}  
(“bright NOUN”)  

Byl jasný studený dubnový den a hodiny odbíjely třináctou

V A A A A N J N V C

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- Is this a good edge?

jasný ← den
(“bright day”)

jasný ← N
(“bright NOUN”)

“IT was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Is this a good edge?

jasný ← den
(“bright day”)

jasný ← N
(“bright NOUN”)

A ← N

Byl  jasný  studený  dubnový  den  a  hodiny  odbíjely  třináctou

V   A   A   A   A   N   J   N   V   C

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- Is this a good edge?

"It was a bright cold day in April and the clocks were striking thirteen"
Edge-Factored Parsers

• How about this competing edge?

“it was a bright cold day in april and the clocks were striking thirteen”
Edge-Factored Parsers

• How about this competing edge?

not as good, lots of red ...

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• How about this competing edge?

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- How about this competing edge?

jasný ← hodiny
(“bright clocks”)

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• How about this competing edge?

**jasný ⇙ hodiny**
(“bright clocks”)
... undertrained ...

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- How about this competing edge?

jasný ← hodiny
(“bright clocks”)
... undertrained ...
Edge-Factored Parsers

- How about this competing edge?

jasný ↔ hodiny
("bright clocks")
... undertrained ...

jasn ↔ hodi
("bright clock,
stems only")

“Byl jasný studený dubnový den a hodiny odbíjely třináctou

V A A A N J N V C

byl jasn stud dubn den a hodi odbí třin

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- How about this competing edge?

jasný ↔ hodiny
("bright clocks")
... undertrained ...

jasn ↔ hodi
("bright clock,”
stems only)

Byl  jasný  studený  dubnový  den  a  hodiny  odbíjely  třináctou
V    A      A      A      N  J  N    V      C

byl  jasn  stud  dubn  den  a  hodi  odbí  třin

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• How about this competing edge?

jasný ← hodiny
("bright clocks")
... undertrained ...

jasn ← hodi
("bright clock," stems only)

A_{singular} ← N_{plural}

Byl jasný studený dubnový den a hodiny odbíjely třináctou

V A A A N J N V C
byl jasn stud dubn den a hodi odbí třin

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- How about this competing edge?

jasný ⇄ hodiny
("bright clocks")
... undertrained ...

jasn ⇄ hodi
("bright clock,"
stems only)

A_{singular} ⇄ N_{plural}

Byl jasný studený dubnový den a hodiny odbíjely třináctou

V A A A N J N V C

byl jasn stud dubn den a hodi odbí třin

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- How about this competing edge?

jasný $\leftarrow$ hodiny

A $\leftarrow$ N

where N follows a conjunction

jasn $\leftarrow$ hodi

(“bright clock,” stems only)

A$_{\text{singular}}$ $\leftarrow$ N$_{\text{plural}}$

Byl jasný studený dubnový den a hodiny odbíjely třináctou

V A A A N J N V C

byl jasn stud dubn den a hodi odbí třin

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

• Which edge is better?
• “bright day” or “bright clocks”?

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- Which edge is better?
- Score of an edge $e = \theta \cdot \text{features}(e)$
- Standard algos $\Rightarrow$ valid parse with max total score

“It was a bright cold day in April and the clocks were striking thirteen”
Edge-Factored Parsers

- Which edge is better?
- Score of an edge \( e = \theta \cdot \text{features}(e) \)
- Standard algos \( \rightarrow \) valid parse with max total score

“It was a bright cold day in April and the clocks were striking thirteen”
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

![Diagram of local factors in a graphical model with nodes labeled 'find', 'preferred', and 'tags' connected by lines.]

```plaintext
... find preferred tags ...
```
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

![Diagram of a graphical model with nodes labeled 'find', 'preferred', and 'tags'. The sentence is shaded as 'Observed input sentence (shaded)'.]
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

Possible tagging (i.e., assignment to remaining variables)

```
... v v v ...
```

```
find preferred tags
```

Observed input sentence (shaded)
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

Possible tagging (i.e., assignment to remaining variables)
Another possible tagging

Observed input sentence (shaded)
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

Possible tagging (i.e., assignment to remaining variables)
Another possible tagging

Observed input sentence (shaded)
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

"Binary" factor that measures compatibility of 2 adjacent tags
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

"Binary" factor that measures compatibility of 2 adjacent tags

Model reuses same parameters at this position
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

"Unary" factor evaluates this tag

```
find
preferred
tags
```

<table>
<thead>
<tr>
<th>v</th>
<th>n</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>0</td>
</tr>
</tbody>
</table>
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

"Unary" factor evaluates this tag
Its values depend on corresponding word

```
find
preferred
```

\[
\begin{array}{c}
\text{v} \quad 0.2 \\
\text{n} \quad 0.2 \\
\text{a} \quad 0
\end{array}
\]
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

"Unary" factor evaluates this tag
Its values depend on corresponding word

- tags can't be adj

```
  v 0.2
  n 0.2
  a 0
```
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

“Unary” factor evaluates this tag
Its values depend on corresponding word

... find preferred tags ...

<table>
<thead>
<tr>
<th>0.2</th>
<th>0.2</th>
<th>0</th>
</tr>
</thead>
</table>

\[ v \]
\[ n \]
\[ a \]
First, a familiar example

- Conditional Random Field (CRF) for POS tagging

"Unary" factor evaluates this tag
Its values depend on corresponding word

(find) preferred tags

(could be made to depend on entire observed sentence)
First, a familiar example
- Conditional Random Field (CRF) for POS tagging
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

"Unary" factor evaluates this tag
Different unary factor at each position

```
  find  preferred  tags
```

```plaintext
v  0.3
n  0.02
a  0

v  0.3
n  0
a  0.1

v  0.2
n  0.2
a  0
```
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

![Diagram of a Conditional Random Field (CRF) for POS tagging with nodes labeled 'find', 'preferred', and 'tags' and edge weights illustrated with tables of v, n, and a values.]

- \( v \): 0.3
- \( n \): 0.02
- \( a \): 0

- \( v \): 0.3
- \( n \): 0
- \( a \): 0.1

- \( v \): 0.2
- \( n \): 0.2
- \( a \): 0
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

\[ p(\text{v a n}) \text{ is proportional to the product of all factors' values on v a n} \]
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

\[ p(van) \text{ is proportional to the product of all factors' values on } van \]
Local factors in a graphical model

- First, a familiar example
  - Conditional Random Field (CRF) for POS tagging

\[ p(v, a, n) \text{ is proportional to the product of all factors' values on } v, a, n \]

\[ \text{find} \quad \text{preferred} \quad \text{tags} \]

\[ 1 \times 3 \times 0.3 \times 0.1 \times 0.2 \ldots \]
Graphical Models for Parsing

- First, a labeling example
  - CRF for POS tagging
- Now let’s do dependency parsing!
  - $O(n^2)$ boolean variables for the possible links

... find preferred links ...
Graphical Models for Parsing

• First, a labeling example
  ✤ CRF for POS tagging

• Now let’s do dependency parsing!
  ✤ $O(n^2)$ boolean variables for the possible links

... find preferred links ...
Graphical Models for Parsing

- First, a labeling example
  - CRF for POS tagging

- Now let’s do dependency parsing!
  - $O(n^2)$ boolean variables for the possible links
Graphical Models for Parsing

- First, a labeling example
  - CRF for POS tagging

- Now let’s do dependency parsing!
  - $O(n^2)$ boolean variables for the possible links

Possible parse...

... find preferred links ...
Graphical Models for Parsing

• First, a labeling example
  ✤ CRF for POS tagging

• Now let’s do dependency parsing!
  ✤ O(n^2) boolean variables for the possible links

Possible parse...

... find preferred links ...

... and variable assignment
Graphical Models for Parsing

• First, a labeling example
  ✤ CRF for POS tagging

• Now let’s do dependency parsing!
  ✤ $O(n^2)$ boolean variables for the possible links

Another parse...

... find preferred links ...
Graphical Models for Parsing

• First, a labeling example
  ✤ CRF for POS tagging

• Now let’s do dependency parsing!
  ✤ $O(n^2)$ boolean variables for the possible links
Graphical Models for Parsing

• First, a labeling example
  ✤ CRF for POS tagging

• Now let’s do dependency parsing!
  ✤ $O(n^2)$ boolean variables for the possible links

An illegal parse...

... find preferred links ...

... with a cycle!
Graphical Models for Parsing

• First, a labeling example
  ✤ CRF for POS tagging

• Now let’s do dependency parsing!
  ✤ $O(n^2)$ boolean variables for the possible links

An illegal parse...

... find preferred links ...

... with a cycle and multiple parents!
Local Factors for Parsing

• What factors determine parse probability?

... find preferred links ...
Local Factors for Parsing

• What factors determine parse probability?

❖ Unary factors to score each link in isolation
Local Factors for Parsing

• What factors determine parse probability?
  ✤ Unary factors to score each link in isolation

... find preferred links ...
Local Factors for Parsing

- What factors determine parse probability?
  - Unary factors to score each link in isolation

- But what if the best assignment isn’t a tree?

... find preferred links ...
Global Factors for Parsing

- What factors determine parse probability?

  - Unary factors to score each link in isolation
Global Factors for Parsing

- What factors determine parse probability?
  - Unary factors to score each link in isolation
  - Global TREE factor to require links to form a legal tree
Global Factors for Parsing

- What factors determine parse probability?
  - Unary factors to score each link in isolation
  - Global TREE factor to require links to form a legal tree
  - A hard constraint: potential is either 0 or 1

... find preferred links ...

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ffffff</td>
<td>0</td>
<td>ffffff</td>
<td>0</td>
<td>ffffff</td>
</tr>
<tr>
<td>ffffff</td>
<td>0</td>
<td>ffffff</td>
<td>0</td>
<td>fffffff</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ffffff</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tttt</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Global Factors for Parsing

• What factors determine parse probability?
  
  ✤ Unary factors to score each link in isolation
  
  ✤ Global TREE factor to require links to form a legal tree
  
  • A *hard constraint*: potential is either 0 or 1

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>...</th>
<th>0</th>
<th>1</th>
<th>...</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ffffffff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fffftt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ffffff</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fftffft</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tttttt</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

we're legal!
Global Factors for Parsing

• What factors determine parse

✓ Unary factors to score each link in isolation
✓ Global TREE factor to require links to form a legal tree

• A hard constraint: potential is either 0 or 1

So far, this is equivalent to edge-factored parsing

We're legal!

64 entries (0/1)

... find preferred links ...
Global Factors for Parsing

- What factors determine parse:
  - Unary factors to score each link in isolation
  - Global TREE factor to require links to form a legal tree
  - A hard constraint: potential is either 0 or 1

So far, this is equivalent to edge-factored parsing

64 entries (0/1)

Note: traditional parsers don’t loop through this table to consider exponentially many trees one at a time. They use combinatorial algorithms; so should we!

... fin
Local Factors for Parsing

• What factors determine parse probability?
  ✤ Unary factors to score each link in isolation
  ✤ Global TREE factor to require links to form a legal tree
    • A hard constraint: potential is either 0 or 1
  ✤ Second order effects: factors on 2 variables
    • Grandparent–parent–child chains

... find preferred links ...
Local Factors for Parsing

- What factors determine parse probability?
  - Unary factors to score each link in isolation
  - Global TREE factor to require links to form a legal tree
    - A *hard constraint*: potential is either 0 or 1
  - Second order effects: factors on 2 variables
    - Grandparent–parent–child chains
Local Factors for Parsing

• What factors determine parse probability?
  ✤ Unary factors to score each link in isolation
  ✤ Global TREE factor to require links to form a legal tree
    • A hard constraint: potential is either 0 or 1
  ✤ Second order effects: factors on 2 variables
    • Grandparent–parent–child chains
    • No crossing links
    • Siblings
  ✤ Hidden morphological tags
  ✤ Word senses and subcategorization frames
Great Ideas in ML: Message Passing

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

there's 1 of me

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

*Count the soldiers*

- 1 of me
- 2 before you
- 3 behind you
- only see my incoming messages

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

there's 1 of me

Belief: Must be 2 + 1 + 3 = 6 of us

2 before you

only see my incoming messages

3 behind you

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

belief: Must be $2 + 1 + 3 = 6$ of us

2 before you

there's only see 1 of me my incoming messages

3 behind you

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

Belief:
Must be
\[ 2 + 1 + 3 = 6 \]
of us

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Count the soldiers

only see my incoming messages

there’s 1 of me

1 before you

4 behind you

Belief:
Must be $1 + 1 + 4 = 6$
of us

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of the tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

*Each soldier receives reports from all branches of tree*

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of the tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

Belief: Must be 14 of us

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of the tree.

Belief: Must be 14 of us.

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

adapted from MacKay (2003) textbook
Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree

Belief: Must be 14 of us

Wouldn’t work correctly with a “loopy” (cyclic) graph

adapted from MacKay (2003) textbook
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = “sum-product algorithm”
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = “sum-product algorithm”

...
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = “sum-product algorithm”
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = "sum-product algorithm"
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = “sum-product algorithm”

\[ \alpha, \beta \]

\[ \begin{array}{c|c|c}
    v & n & a \\
    \hline
    7 & 2 & 1 \\
    \hline
    0 & 2 & 1 \\
    2 & 1 & 0 \\
    0 & 3 & 1 \\
  \end{array} \]

\[ \begin{array}{c|c|c|c}
        & v & n & a \\
  \hline
  \alpha & 3 & 1 & 6 \\
  \beta & 0 & 2 & 1 \\
        & 2 & 1 & 0 \\
        & 0 & 3 & 1 \\
\end{array} \]

belief

\[ \begin{array}{c}
    v \\
    n \\
    a \\
  \end{array} \]

\[ \begin{array}{c}
    1. \\
    \hat{0} \\
    4. \\
  \end{array} \]

message

\[ \begin{array}{c|c|c}
    v & n & a \\
    \hline
    0.3 & 0 & 0.1 \\
  \end{array} \]

find

preferred

tags
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = “sum-product algorithm”

In the CRF, message passing = forward-backward = “sum-product algorithm”
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = “sum-product algorithm”

![Diagram showing message passing in a CRF with nodes and messages labeled with values like α, β, v, n, a, and belief values such as 0.3, 0, and 0.1. The diagram illustrates the forward and backward messages with arrows between nodes. The diagram also includes labels find, preferred, and tags.]
In the CRF, message passing = forward-backward = “sum-product algorithm”
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = “sum-product algorithm”

In the CRF, message passing = forward-backward = “sum-product algorithm”
Great ideas in ML: Forward-Backward

- In the CRF, message passing = forward-backward = “sum-product algorithm”
Sum-Product Equations

- Message from variable $v$ to factor $f$

$$m_{v \rightarrow f}(x) = \prod_{f' \in N(v) \setminus \{f\}} m_{f' \rightarrow v}(x)$$

- Message from factor $f$ to variable $v$
Recipe for Conditional Training of $p(y \mid x)$

1. Gather constraints/features from training data
   \[ \alpha_{iy} = \tilde{E}[f_{iy}] = \sum_{x_j, y_j \in D} f_{iy}(x_j, y_j) \]

2. Initialize all parameters to zero

3. Classify training data with current parameters; calculate expectations
   \[ E_\Theta[f_{iy}] = \sum_{x_j \in D} \sum_{y^{'}} p_\Theta(y^{'} \mid x_j) f_{iy}(x_j, y^{'}) \]

4. Gradient is
   \[ \tilde{E}[f_{iy}] - E_\Theta[f_{iy}] \]

5. Take a step in the direction of the gradient

6. Repeat from 3 until convergence
Great ideas in ML: Forward-Backward

\[
\begin{array}{c|c|c}
\alpha & v & 3 \\
\hline
n & 1 \\
\hline
a & 6 \\
\end{array}
\quad
\begin{array}{c|c|c}
\beta & v & 2 \\
\hline
n & 1 \\
\hline
a & 7 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c|c|c}
& v & n & a & 0.3 & 0 & 0.1 \\
\hline
v & 3 \\
\hline
n & 1 \\
\hline
a & 6 \\
\end{array}
\quad
\begin{array}{c|c|c|c|c|c|c}
& v & n & a & 2 & 1 & 7 \\
\hline
v & 2 \\
\hline
n & 1 \\
\hline
a & 7 \\
\end{array}
\]

find

preferred

tags
Great ideas in ML: Forward-Backward

\[ \alpha \]

\[ \beta \]

\[ v \quad n \quad a \]

\[ 3 \quad 1 \quad 6 \]

\[ v \quad n \quad a \]

\[ 2 \quad 1 \quad 7 \]

\[ v \quad n \quad a \]

\[ 0.3 \quad 0 \quad 0.1 \]

find

preferred

tags
Great ideas in ML: Forward-Backward

- Extend CRF to “skip chain” to capture non-local factor
Great ideas in ML: Forward-Backward

- Extend CRF to “skip chain” to capture non-local factor
  - More influences on belief 😊

```
 n a 6
 v 3
 n 1
 a 6
 v 3
 n 1
 a 6
 v 5.4
 n 0 a 25.2
 v 0.3
 n 0 a 0.1

α
β
```

find preferred tags
Great ideas in ML: Forward-Backward

- Extend CRF to “skip chain” to capture non-local factor
  - More influences on belief 😊
  - Graph becomes loopy 😞
Great ideas in ML: Forward-Backward

- Extend CRF to “skip chain” to capture non-local factor
  - More influences on belief 😊
  - Graph becomes loopy 😞

Red messages not independent? Pretend they are!
Great ideas in ML: Forward-Backward

- Extend CRF to “skip chain” to capture non-local factor
  - More influences on belief 😊
  - Graph becomes loopy 😞

Red messages not independent? Pretend they are! “Loopy Belief Propagation”
Terminological Clarification

propagation
Terminological Clarification

belief propagation
Terminological Clarification

loopy belief propagation
Terminological Clarification

\[
\text{loopy belief} \rightarrow \text{propagation}
\]
Terminological Clarification

(loopy belief) propagation
Terminological Clarification

\[
\{ \text{loopy} \rightarrow \text{belief} \rightarrow \text{propagation} \}
\]
Terminological Clarification

\(\text{loopy} \quad \text{belief} \quad \text{propagation}\)

\(\text{loopy} \quad \text{belief propagation}\)
Propagating Global Factors

• Loopy belief propagation is easy for local factors

• How do combinatorial factors (like TREE) compute the message to the link in question?

✦ “Does the TREE factor think the link is probably t given the messages it receives from all the other links?”

... find preferred links ...
Propagating Global Factors

• Loopy belief propagation is easy for local factors

• How do combinatorial factors (like TREE) compute the message to the link in question?
  
  “Does the TREE factor think the link is probably \( t \) given the messages it receives from all the other links?”
Propagating Global Factors

- Loopy belief propagation is easy for local factors.
- How do combinatorial factors (like TREE) compute the message to the link in question?

“Does the TREE factor think the link is probably t given the messages it receives from all the other links?”

<table>
<thead>
<tr>
<th>TREE factor</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ffffff</td>
<td>0</td>
</tr>
<tr>
<td>ffffft</td>
<td>0</td>
</tr>
<tr>
<td>fffftf</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>fftfttt</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>tttttt</td>
<td>0</td>
</tr>
</tbody>
</table>
Propagating Global Factors

• How does the TREE factor compute the message to the link in question?

❖ “Does the TREE factor think the link is probably t given the messages it receives from all the other links?”

<table>
<thead>
<tr>
<th>TREE factor</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ffffffff</td>
<td>0</td>
</tr>
<tr>
<td>ffffff</td>
<td>0</td>
</tr>
<tr>
<td>ffffft</td>
<td>0</td>
</tr>
<tr>
<td>fffttf</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>fftttf</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>tttttt</td>
<td>0</td>
</tr>
</tbody>
</table>
Propagating Global Factors

• How does the TREE factor compute the message to the link in question?
  ✤ “Does the TREE factor think the link is probably t given the messages it receives from all the other links?”

Old-school parsing to the rescue!

This is the outside probability of the link in an edge-factored parser!

∴ TREE factor computes all outgoing messages at once (given all incoming messages)

Projective case: total $O(n^3)$ time by inside-outside

Non-projective: total $O(n^3)$ time by inverting Kirchhoff matrix
Graph Theory to the Rescue!

Tutte’s **Matrix-Tree Theorem** (1948)

The **determinant** of the Kirchoff (aka Laplacian) adjaceny matrix of directed graph $G$ without row and column $r$ is equal to the **sum of scores of all directed spanning trees** of $G$ rooted at node $r$. 
Tutte’s Matrix-Tree Theorem (1948)

The determinant of the Kirchoff (aka Laplacian) adjacency matrix of directed graph $G$ without row and column $r$ is equal to the sum of scores of all directed spanning trees of $G$ rooted at node $r$.

Exactly the $Z$ we need!
Graph Theory to the Rescue!

Tutte's Matrix-Tree Theorem (1948)

The determinant of the Kirchoff (aka Laplacian) adjacency matrix of directed graph $G$ without row and column $r$ is equal to the sum of scores of all directed spanning trees of $G$ rooted at node $r$.

Exactly the $Z$ we need!

$O(n^3)$ time!
Kirchoff (Laplacian) Matrix

\[
\begin{bmatrix}
0 & -s(1,0) & -s(2,0) & \ldots & -s(n,0) \\
0 & 0 & -s(2,1) & \ldots & -s(n,1) \\
0 & -s(1,2) & 0 & \ldots & -s(n,2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & -s(1,n) & -s(2,n) & \ldots & 0
\end{bmatrix}
\]

- Negate edge scores
- Sum columns (children)
- Strike root row/col.
- Take determinant
Kirchhoff (Laplacian) Matrix

$$
\begin{bmatrix}
0 & -s(1,0) & -s(2,0) & \ldots & -s(n,0) \\
0 & 0 & -s(2,1) & \ldots & -s(n,1) \\
0 & -s(1,2) & 0 & \ldots & -s(n,2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & -s(1,n) & -s(2,n) & \ldots & 0
\end{bmatrix}
$$

- Negate edge scores
- Sum columns (children)
- Strike root row/col.
- Take determinant
Kirchoff (Laplacian) Matrix

\[
\begin{bmatrix}
0 & -s(1,0) & -s(2,0) & \cdots & -s(n,0) \\
0 & \sum_{j \neq 1} s(1,j) & -s(2,1) & \cdots & -s(n,1) \\
0 & -s(1,2) & \sum_{j \neq 2} s(2,j) & \cdots & -s(n,2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & -s(1,n) & -s(2,n) & \cdots & \sum_{j \neq n} s(n,j)
\end{bmatrix}
\]

- Negate edge scores
- Sum columns (children)
- Strike root row/col.
- Take determinant
Kirchoff (Laplacian) Matrix

\[
\begin{pmatrix}
\sum_{j \neq 1} s(1,j) & -s(2,1) & \cdots & -s(n,1) \\
-s(1,2) & \sum_{j \neq 2} s(2,j) & \cdots & -s(n,2) \\
\vdots & \vdots & \ddots & \vdots \\
-s(1,n) & -s(2,n) & \cdots & \sum_{j \neq n} s(n,j)
\end{pmatrix}
\]

- Negate edge scores
- Sum columns (children)
- Strike root row/col.
- Take determinant
### Kirchoff (Laplacian) Matrix

\[
\begin{array}{cccc}
\sum_{j \neq 1} s(1, j) & -s(2,1) & \cdots & -s(n,1) \\
-s(1,2) & \sum_{j \neq 2} s(2, j) & \cdots & -s(n,2) \\
\vdots & \vdots & \ddots & \vdots \\
-s(1,n) & -s(2,n) & \cdots & \sum_{j \neq n} s(n,j) \\
\end{array}
\]

- Negate edge scores
- Sum columns (children)
- Strike root row/col.
- Take determinant

N.B.: This allows multiple children of root, but see Koo et al. 2007.
Transition-Based Parsing

- Linear time
- Online
- Train a classifier to predict next action
- Deterministic or beam-search strategies
- But... generally less accurate
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Start state: \( ([], [1, \ldots, n], \{ \}) \)

Final state: \( (S, [], A) \)

Shift: \( (S, i | B, A) \Rightarrow (S | i, B, A) \)

Reduce: \( (S | i, B, A) \Rightarrow (S, B, A) \)

Right-Arc: \( (S | i, j | B, A) \Rightarrow (S | i | j, B, A \cup \{i \rightarrow j\}) \)

Left-Arc: \( (S | i, j | B, A) \Rightarrow (S, j | B, A \cup \{i \leftarrow j\}) \)
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack  Buffer  Arcs
[ ]s  [who, did, you, see]B  {}

Diagram:

Stack: [ ]s  Buffer: [who, did, you, see]B  Arcs: {}

Diagram nodes:
- OBJ: who
- SBJ: did
- VG: you
- see
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack | Buffer | Arcs
---|---|---
[who]_S | [did, you, see]_B | {}
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack | Buffer | Arcs

\[ ]_S | [did, you, see]_B | \{ who \leftarrow^{OBJ} \text{ did} \}

Left-arc
OBJ

[Diagram of dependency parsing with arcs and labels: OBJ, SBJ, VG, who, did, you, see]
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack
[did]\textsubscript{S}

Buffer
[you, see]\textsubscript{B}

Arcs
\{ who \textsubscript{OBJ} \rightleftharpoons \text{did} \}

Shift

\textbf{OBJ} \quad \textbf{SBJ} \quad \text{VG}

\begin{itemize}
  \item who \quad \text{did} \quad \text{you} \quad \text{see}
\end{itemize}
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

**Stack**

[did, you]$_S$

**Buffer**

[see]$_B$

**Arcs**

{ who $\xrightarrow{OBJ}$ did, 
did $\xrightarrow{SBJ}$ you }

*Right-arc*  

*SBJ*
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack

[did]_S

Buffer

[see]_B

Arcs

\{ who \overset{OBJ}{\leftarrow} \text{did,} \\
\text{did} \overset{SBJ}{\rightarrow} \text{you} \}

Reduce

who \quad \text{did} \quad \text{you} \quad \text{see}
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack
[did, see]_S

Buffer
[ ]_B

Arcs
{ who OBJ did, 
did SBJ you, 
did VG see }
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack  Buffer  Arrows
[did, you]_S  [see]_B  { who \(\xrightarrow{\text{OBJ}}\) did,  
did \(\xrightarrow{\text{SBJ}}\) you }

Right-arc  SBJ
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack

[did, you]$_S$

Buffer

[see]$_B$

Arcs

\{ who $\xleftarrow{\text{OBJ}}$ did, \\
    did $\xrightarrow{\text{SBJ}}$ you \}

Right-arc

SBJ

Choose action w/best classifier score

100k - 1M features
Transition-Based Parsing

Arc-eager shift-reduce parsing (Nivre, 2003)

Stack
[did, you]$_S$ [s]

Bare-Bones Dependency Parsing

Very fast linear-time performance
WSJ 23 (2k sentences) in 3 s

did $\rightarrow$ you}

Right-arc
SBJ

Choose action w/best classifier score
100k - 1M features

Right-arc
SBJ

Choose action w/best classifier score
100k - 1M features