Analyzing Errors of Unsupervised Learning

ACL 2008    Columbus, Ohio

June 18, 2008

Percy Liang    Dan Klein
Unsupervised grammar induction

Goal: induce hidden syntax

The man ate a tasty sandwich
Unsupervised grammar induction

Goal: induce hidden syntax

DT—NN—VBD—DT—JJ—NN
The man ate a tasty sandwich

POS tagging
Unsupervised grammar induction

**Goal**: induce hidden syntax

```
DT NN VBD DT JJ NN
The man ate a tasty sandwich
```

POS tagging

```
DT NN VBD
The man ate a tasty sandwich
```

POS tagging

```
DT
```

POS tagging

```
JJ NN
```

POS tagging
Unsupervised grammar induction

Goal: induce hidden syntax

POS tagging

Constituency parsing
Unsupervised grammar induction

**Goal:** induce hidden syntax

The man ate a tasty sandwich

POS tagging

Constituency parsing

For example, on POS tagging using HMMs:

Unsupervised using EM \(\approx 60\%\)
Unsupervised grammar induction

Goal: induce hidden syntax

The man ate a tasty sandwich

POS tagging

Constituency parsing

For example, on POS tagging using HMMs:

Unsupervised using EM ≈ 60%

Supervised ≥ 90%
Unsupervised grammar induction

Goal: induce hidden syntax

The man ate a tasty sandwich

POS tagging

Constituency parsing

For example, on POS tagging using HMMs:

Unsupervised using EM ≈ 60%
Supervised ≥ 90%

Why does EM fail?
Four types of errors:
Four types of errors:

- Optimization error
- Local optima
Four types of errors:

Optimization error
  Local optima

Estimation error
  Limited data
Four types of errors:

Optimization error
  Local optima

Estimation error
  Limited data

Approximation error
  Likelihood objective $\nRightarrow$ accuracy
Four types of errors:

- **Optimization error**
  - Local optima

- **Estimation error**
  - Limited data

- **Approximation error**
  - Likelihood objective $\not\equiv$ accuracy

- **Identifiability error**
  - Different parameter settings $\rightarrow$ same objective
Approximation error

Problem: model likelihood $\not\Leftrightarrow$ prediction accuracy
Approximation error

Problem: model likelihood $\not\approx$ prediction accuracy

PCFG (EM starting from supervised parameter estimate):
Approximation error

Problem: model likelihood $\nRightarrow$ prediction accuracy

PCFG (EM starting from supervised parameter estimate):

What qualitative changes is EM making?
Migrations

For the HMM:

**Truth**
DT NN NN RB VBD NNS
The chief executive allegedly made contributions

**Iteration 1**
DT JJ NN RB VBN NNS
The chief executive allegedly made contributions
Migrations

For the HMM:

Truth
DT NN NN RB VBD NNS
The chief executive allegedly made contributions

Iteration 1
DT JJ NN RB VBN NNS
The chief executive allegedly made contributions

Summarize changes by a set of migrations:

NN VBD
JJ VBN

5
Migrations

For the HMM:

Truth

The chief executive allegedly made contributions

Iteration 1

The chief executive allegedly made contributions

Summarize changes by a set of migrations:

NN → NN
JJ → NN

VBD → made
VBN → made
Migrations

For the HMM:

Truth
DT NN NN RB VBD NNS
The chief executive allegedly made contributions

Iteration 1
DT JJ NN RB VBN NNS
The chief executive allegedly made contributions

Summarize changes by a set of migrations:

NN→NN  VBD→made
JJ→NN  VBN→made

What are the prominent migrations over the entire corpora?
Sentence-initial nouns are often proper

\[ \text{Iteration 1} \]

\text{START} \rightarrow \text{NNN} \hspace{1cm} \text{Revenue/NN/NNP rose}
## Top HMM migrations

### Iteration 1

<table>
<thead>
<tr>
<th>START  →</th>
<th>NN</th>
<th>NNP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sentence-initial nouns are often proper

\[
\text{START Revenue/NN/NNP rose}
\]

<table>
<thead>
<tr>
<th>NN  →</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ</td>
<td></td>
</tr>
</tbody>
</table>

Noun adjuncts → adjectives (inconsistent gold tags)

\[
\text{chief/NN/JJ executive/NN officer}
\]
Sentence-initial nouns are often proper

\[\text{START Revenue/NN/NNP} \ \text{rose}\]

Noun adjuncts $\rightarrow$ adjectives (inconsistent gold tags)

\[\text{chief/NN/JJ executive/NN officer}\]

Inconsistent gold tags

\[\text{UBS Securities/NNP/NNPS}\]
Top HMM migrations

Iteration 1

START $\rightarrow$ NNP

Sentence-initial nouns are often proper

START Revenue/NN/NNP rose

NN $\rightarrow$ NN

Noun adjuncts $\rightarrow$ adjectives (inconsistent gold tags)

chief/NN/JJ executive/NN officer

NNP $\rightarrow$ NNPS

Inconsistent gold tags

UBS Securities/NNP/NNPS

Iteration 2

NN $\rightarrow$ NN

(same as above)

JJ $\rightarrow$ NN

(same as above)

START $\rightarrow$ NNP

(same as above)

JJ $\rightarrow$ TO

Inconsistent gold tags

contribute much/JJ/RB to
Meta-modeling for PCFGs

Truth
There is an element of make-work

Iteration 1
There is an element of make-work
Meta-modeling for PCFGs

Truth

There is an element of make-work

Iteration 1

There is an element of make-work
Meta-modeling for PCFGs

There is an element of make-work

Migrations less clear due to uncertainty in tree structure...
Meta-modeling for PCFGs

Migrations less clear due to uncertainty in tree structure...
Meta-modeling for PCFGs

Migrations less clear due to uncertainty in tree structure...

Our approach: use a meta-model
- Migrations are hidden alignments to be learned
- Fit using EM
Meta-modeling for PCFGs

Migrations less clear due to uncertainty in tree structure...

Our approach: use a meta-model
- Migrations are hidden alignments to be learned
- Fit using EM (convex, similar to IBM model 1)
Top PCFG migrations learned by meta-model

**Iteration 1**

```
S
  / \  \
 RB   VP
  / \  \
 RB   VP
```

Sentential adverbs $\rightarrow$ VP adverbs
Top PCFG migrations learned by meta-model

Iteration 1

S

RB VP

RB VP

VP

NP PP

NP

VP PP

VP

PPs raised from NPs to verbal level

Sentential adverbs $\rightarrow$ VP adverbs
Top PCFG migrations learned by meta-model

**Iteration 1**

```
S
  +---+---+
  | RB | VP |
  +-----+
    | VP |
    +-----+
      | NP |
      | PP |
      +----+
        | VP |
        +----+
```

Sentential adverbs → VP adverbs

PPs raised from NPs to verbal level

**Iteration 2**

```
NNP
  +---+---+
  | NP | NP |
  +-----+
    | NNP |
```

Right-branching → left-branching structures
Top PCFG migrations learned by meta-model

**Iteration 1**

- **Sentential adverbs → VP adverbs**
- **PPs raised from NPs to verbal level**

**Iteration 2**

- **Right-branching → left-branching structures**
- **PP raised to higher VP**
Meta-modeling summary

• Meta-model: a diagnostic tool to analyze errors systematically
Meta-modeling summary

• Meta-model: a diagnostic tool to analyze errors systematically

• General phenomenon: regularization of syntactic structure
Meta-modeling summary

- Meta-model: a diagnostic tool to analyze errors systematically

- General phenomenon: **regularization** of syntactic structure

  ✓ Approximation error
  Identifiability error
  Estimation error
  Optimization error
Identifiability error

\( x \): input sentence

\( y \): hidden output

\( p_\theta(x, y) \): joint distribution with parameters \( \theta \)
Identifiability error

\( x \): input sentence

\( y \): hidden output

\( p_{\theta}(x, y) \): joint distribution with parameters \( \theta \)

Non-identifiability:

\[
\begin{array}{c}
\theta_1 \\
? \\
\theta_2
\end{array}
\]
Identifiability error

\textbf{x}: input sentence

\textbf{y}: hidden output

\( p_\theta(x, y) \): joint distribution with parameters \( \theta \)

\textbf{Non-identifiability:}

Learning is indifferent... \( p_{\theta_1}(x) = p_{\theta_2}(x) \)

\( \theta_1 \) ? \( \theta_2 \)
Identifiability error

\(x\): input sentence

\(y\): hidden output

\(p_\theta(x, y)\): joint distribution with parameters \(\theta\)

Non-identifiability:

Learning is indifferent...

\[ p_{\theta_1}(x) = p_{\theta_2}(x) \]

\[ \theta_1 \neq \theta_2 \]

but matters to prediction (bad!)

\[ p_{\theta_1}(y \mid x) \neq p_{\theta_2}(y \mid x) \]
Examples of non-identifiability

- Label symmetries
  
  \[
  \begin{array}{c}
  1 \quad 2 \\
  \downarrow \quad \downarrow \\
  a \quad b
  \end{array}
  \quad \text{and} \quad
  \begin{array}{c}
  2 \quad 1 \\
  \downarrow \quad \downarrow \\
  a \quad b
  \end{array}
  \]
  both generate abababab...
Examples of non-identifiability

- Label symmetries
  
  \[ \begin{align*}
  1 & \rightarrow 2 \\
  \downarrow & \quad \downarrow \\
  a & \quad b \\
  \end{align*} \quad \text{and} \quad \begin{align*}
  2 & \rightarrow 1 \\
  \downarrow & \quad \downarrow \\
  a & \quad b \\
  \end{align*} \]
  both generate abababab...

- \( K \)-state HMM (if true distribution is \(< K\)-state HMM)

  \[ \begin{align*}
  1 & \rightarrow 2 \\
  \downarrow & \quad \downarrow \\
  a & \quad b \\
  \end{align*} \quad \text{and} \quad \begin{align*}
  1 & \rightarrow 2 \rightarrow 3 \\
  \downarrow & \quad \downarrow \quad \downarrow \\
  a & \quad b & \quad a \\
  \end{align*} \]
  both generate abababab...
Examples of non-identifiability

- **Label symmetries**
  
  \[
  \begin{array}{c}
  1 \quad 2 \\
  \downarrow & \downarrow
  \end{array}
  \quad \text{and} \quad 
  \begin{array}{c}
  2 \quad 1 \\
  \downarrow & \downarrow
  \end{array}
  \text{ both generate abababab...}
  
  \begin{array}{cc}
  a & b \\
  \downarrow & \downarrow
  \end{array}
  \quad \begin{array}{cc}
  a & b \\
  \downarrow & \downarrow
  \end{array}
  
- **$K$-state HMM (if true distribution is < $K$-state HMM)**
  
  \[
  \begin{array}{c}
  1 \quad 2 \\
  \downarrow & \downarrow
  \end{array}
  \quad \text{and} \quad 
  \begin{array}{c}
  1 \quad 2 \quad 3 \\
  \downarrow & \downarrow & \downarrow
  \end{array}
  \text{ both generate abababab...}
  
  \begin{array}{cc}
  a & b \\
  \downarrow & \downarrow
  \end{array}
  \quad \begin{array}{ccc}
  a & b & a \\
  \downarrow & \downarrow & \downarrow
  \end{array}
  
- **PCFG (if true distribution is HMM)**
  
  \[
  \text{and} \quad 
  \text{can both simulate the HMM}
  
  \[
  \text{and} \quad 
  \text{can both simulate the HMM}
Examples of non-identifiability

- Label symmetries
  
  \[
  \begin{array}{ccc}
  1 & 2 \\
  \downarrow & \downarrow \\
  a & b
  \end{array} \quad \text{and} \quad
  \begin{array}{ccc}
  2 & 1 \\
  \downarrow & \downarrow \\
  a & b
  \end{array}
  \]
  both generate abababab...

- $K$-state HMM (if true distribution is $< K$-state HMM)
  
  \[
  \begin{array}{ccc}
  1 & 2 \\
  \downarrow & \downarrow \\
  a & b
  \end{array} \quad \text{and} \quad
  \begin{array}{ccc}
  1 & 2 & 3 \\
  \downarrow & \downarrow & \downarrow \\
  a & b & a
  \end{array}
  \]
  both generate abababab...

- PCFG (if true distribution is HMM)
  
  \[
  \quad \text{and} \quad
  \quad \text{can both simulate the HMM}
  \]

Real data is complex, so last two are not an issue
Identifiability and distance

Given $\theta_1$ and $\theta_2$, how to measure distance between them?

Want distance

$$\begin{pmatrix}
1 & 2 \\
\downarrow & \downarrow \\
a & b
\end{pmatrix},\begin{pmatrix}
2 & 1 \\
\downarrow & \downarrow \\
a & b
\end{pmatrix} = 0$$
Identifiability and distance

Given $\theta_1$ and $\theta_2$, how to measure distance between them?

Want distance $\left( \begin{array}{cc}
1 & 2 \\
a & b \\
\end{array} ,
\begin{array}{cc}
2 & 1 \\
a & b \\
\end{array} \right) = 0$

- Computing label-permutation invariant distance is NP-hard
- We use bipartite matching to find lower and upper bounds
Identifiability and distance

Given $\theta_1$ and $\theta_2$, how to measure distance between them?

Want distance

$$\begin{pmatrix}
  1 & 2 \\
  a & b \\
\end{pmatrix}
\begin{pmatrix}
  2 & 1 \\
  a & b \\
\end{pmatrix} = 0$$

- Computing label-permutation invariant distance is NP-hard
- We use bipartite matching to find lower and upper bounds

- Approximation error
- Identifiability error
- Estimation error
- Optimization error
Estimation and optimization errors

Experiment setup:

- Take some parameters $\theta^*$ (say, supervised estimate on real data)
- Use $\theta^*$ to generate synthetic data
- Can we recover $\theta^*$ using EM?
Estimation and optimization errors

Experiment setup:
- Take some parameters $\theta^*$ (say, supervised estimate on real data)
- Use $\theta^*$ to generate synthetic data
- Can we recover $\theta^*$ using EM?

$\theta^*$ estimation error → global optimum ← optimization error → EM solution
Estimation and optimization errors

Experiment setup:

- Take some parameters $\theta^*$ (say, supervised estimate on real data)
- Use $\theta^*$ to generate synthetic data
- Can we recover $\theta^*$ using EM? No?

HMM on 5K examples:

Accuracy

Sup. init.    Unif. init.

iteration
Estimation and optimization errors

Experiment setup:

- Take some parameters $\theta^*$ (say, supervised estimate on real data)
- Use $\theta^*$ to generate synthetic data
- Can we recover $\theta^*$ using EM? Yes!

HMM on 500K examples:
Optimization error decreases with more data

On HMM model (similar for PCFG and a dependency model):

- Sup. init
- Unif. init

Accuracy vs. number of examples for Synthetic data.
Optimization error decreases with more data

On HMM model (similar for PCFG and a dependency model):

- **Sup. init**
- **Unif. init**

![Accuracy vs. # examples](chart1.png)

- **Synthetic data**
- **Real data**
Optimization error decreases with more data

On HMM model (similar for PCFG and a dependency model):

- **Sup. init**
- **Unif. init**

Why does this phenomenon happen?
- **Intuition**: with more data, EM can pick up the salient patterns more easily
- **Was also shown for mixture of Gaussians** [Srebro, 2006]
Summary

✓ Approximation error

Meta-model: tool for systematic error analysis
Summary

✓ Approximation error
  Meta-model: tool for systematic error analysis

✓ Identifiability error
  Distance robust to label symmetries
Summary

✓ Approximation error
  Meta-model: tool for systematic error analysis

✓ Identifiability error
  Distance robust to label symmetries

✓ Estimation error
  Decreases with more data
Summary

✓ Approximation error
   Meta-model: tool for systematic error analysis

✓ Identifiability error
   Distance robust to label symmetries

✓ Estimation error
   Decreases with more data

✓ Optimization error
   Decreases with more data!