Automatically Learning Measures of Child Language Development

Sam Sahakian and Benjamin Snyder
How does this happen?
When does this happen?

6 months
2 years
4 years
Language Development Metrics

- **MLU (Mean Length of Utterance)** [Brown ’73]
- **Parse depth** [Yngve, ’60]
- **D-Level** [Rosenberg et al., ’87; Covington et al., ’06] [Lu, ’09]
- **IPSYN** [Scarborough, 1990] [Sagae, ’05]
Language Development Metrics

- Drawbacks of previous metrics:
  - Coarse and ad-hoc
  - Questionable validity
  - Accuracy degrades with age

**Question 1:** Can we induce a more accurate metric using statistical learning methods?
Skill as function of time

 Skill acquisition follows sigmoidal curve [Hodgetts '91]

\[ s = \frac{1}{1 + e^{-at+b}} \]
Time as Ground Truth

\[ t = b - \ln \left( \frac{1}{s} - 1 \right) \]

- Invert sigmoid
- Skill as combination of features
- Evaluate learned metric via age prediction error

\[ s \approx \beta \cdot x \]

skill parameters features
Age Prediction Model

\[ t = a(\beta \cdot x) + b \]

- Age window at linear part of sigmoid
- Predict age as linear function of skill
Features

- Pre-defined metrics:
  - MLU
  - Parse Depth
  - D-Level

- Novel features
  - Preposition counts
  - “Be” verb counts
  - Article counts
  - Word frequency
  - Function to content word ratio
Data

- Child speech from transcribed conversations in CHILDES database [MacWhinney, '00]

- Longitudinal studies of 7 children

![Graph showing number of utterances across ages for 7 children.](image)

- Learn via linear regression -- Separately for each child.
Results (lower is better)

<table>
<thead>
<tr>
<th></th>
<th>D-Level</th>
<th>Depth</th>
<th>MLU</th>
<th>All Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>63.795</td>
<td>66.327</td>
<td>64.578</td>
<td>54.041</td>
</tr>
</tbody>
</table>

Mean squared error of age prediction in months
### Results
*(lower is better)*

<table>
<thead>
<tr>
<th></th>
<th>D-Level</th>
<th>Depth</th>
<th>MLU</th>
<th>All Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td>14.037</td>
<td>14.149</td>
<td>11.128</td>
<td>14.175</td>
</tr>
<tr>
<td>Abe</td>
<td>34.69</td>
<td>44.701</td>
<td>34.509</td>
<td>39.931</td>
</tr>
<tr>
<td>Ross</td>
<td>329.64</td>
<td>336.612</td>
<td>345.046</td>
<td>244.071</td>
</tr>
<tr>
<td>Peter</td>
<td>23.58</td>
<td>13.045</td>
<td>8.245</td>
<td>24.128</td>
</tr>
<tr>
<td>Naomi</td>
<td>24.458</td>
<td>28.426</td>
<td>34.956</td>
<td>45.036</td>
</tr>
<tr>
<td>Sarah</td>
<td>12.503</td>
<td>20.878</td>
<td>13.905</td>
<td>6.989</td>
</tr>
<tr>
<td>Nina</td>
<td>7.654</td>
<td>6.477</td>
<td>4.255</td>
<td>3.96</td>
</tr>
<tr>
<td>Mean</td>
<td>63.795</td>
<td>66.327</td>
<td>64.578</td>
<td>54.041</td>
</tr>
</tbody>
</table>

Mean squared error of age prediction in months
Results
(lower is better)

<table>
<thead>
<tr>
<th></th>
<th>D-Level</th>
<th>Depth</th>
<th>MLU</th>
<th>All Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam</td>
<td>14.037</td>
<td>14.149</td>
<td>11.128</td>
<td>14.175</td>
</tr>
<tr>
<td>Abe</td>
<td>34.69</td>
<td>44.701</td>
<td>34.509</td>
<td>39.931</td>
</tr>
<tr>
<td>Ross</td>
<td>329.64</td>
<td>336.612</td>
<td>345.046</td>
<td>244.071</td>
</tr>
<tr>
<td>Peter</td>
<td>23.58</td>
<td>13.045</td>
<td>8.245</td>
<td>24.128</td>
</tr>
<tr>
<td>Naomi</td>
<td>24.458</td>
<td>28.426</td>
<td>34.956</td>
<td>45.036</td>
</tr>
<tr>
<td>Sarah</td>
<td>12.503</td>
<td>20.878</td>
<td>13.905</td>
<td>6.989</td>
</tr>
<tr>
<td>Nina</td>
<td>7.654</td>
<td>6.477</td>
<td>4.255</td>
<td>3.96</td>
</tr>
<tr>
<td>Mean</td>
<td>63.795</td>
<td>66.327</td>
<td>64.578</td>
<td>54.041</td>
</tr>
</tbody>
</table>

Mean squared error of age prediction in months
**Question 2:** Can we learn a metric that generalizes across children?

- **Task:** Train on a set of children, evaluate on a held-out child.
- **Children learn at different rates, so must predict relative mastery, not absolute age.**
Ordering Model

Each iteration trains on 6 children, tests on held-out child

Score each sample as weighted combination of features and feature pairs

Rank speech samples in order of ascending score

\[ y = \sum_{i} \beta_i x_i + \sum_{i,j} \gamma_{ij} x_i x_j \]

Score used for ranking

Sum over features

Sum over feature pairs
Evaluation: Kendall’s $\tau$

\[
\tau = \frac{(\text{num. concordant pairs}) - (\text{num. discordant pairs})}{\frac{1}{2} n(n - 1)}
\]

- Kendall’s rank correlation coefficient
  - Measures similarity between 2 orderings over a set
  - Identical orderings yield +1, independent orderings yield 0
Parameter Estimation

\((\beta^*, \gamma^*) = \argmax_{\beta, \gamma} \sum_{k \in \text{kids}} \tau(k, \beta, \gamma)\)

- \(\tau(k, \beta, \gamma) \equiv \text{Kendall } \tau \text{ between model ordering and true chronological order for child } k\).

- Find best parameters via Nelder-Mead [Nelder and Mead, '65]
  - Gradient-free hill climbing search that shifts parameter values until reaching a local optimum.
Results
(higher is better)

<table>
<thead>
<tr>
<th>MLU</th>
<th>All Features</th>
<th>MLU &amp; Fn. / Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7456</td>
<td>0.7457</td>
<td>0.7780</td>
</tr>
</tbody>
</table>

Average Kendall $\tau$ of model orderings versus true chronological orderings.
Contributions

- New method of inducing language development metrics
- Methodology for validating these metrics
- Increased performance over hand-crafted baseline metrics