With advent of Big-Data (and faster more powerful hardware), the received wisdom is that more (so-so labeled) data = better models. More “silver” labeled data ≥ small expert-annotated “gold” labeled data. More crowdsourced lay-person labels = small expert-annotated labels. More labeled data by mediocre taggers = less labeled data by SotA taggers.

McClosky et al. (2006), Foster et al. (2007), Petrov et al. (2010), ...

Research questions
- Does “more data = better” / “better labeling insignificant at scale” apply to:
  - Semantic role prediction? (SRL “lite”: given a verb and arg, what’s its role?).
  - Role filling / word prediction? (given predicate and role).
  - Thematic fit estimation?
  - Especially as a related task the model was not directly optimized for?

Lexical Resource Improvements
- Relabelling:
  - Original (v1) Replaced with (v2)
    - NLTK/WordNet (Bird et al., 2009) Morfette (Chrupała et al. 2008)
    - MaltParser (Nivre et al. 2006) SpaCy (Honnibal and Montani, 2017)
    - SENNA (Collobert et al., 2011) LSGN (He et al., 2018)
- Extensive work aligning of tokenization schemas over 78M sentences.

Quality vs. quantity
- Better span/role prediction in LSGN vs. SENNA.
- 20% more frame quantity with LSGN.
- Better parsing quality (spaCy vs. MaltParser) and lemmatization (Morphette).

Data quantity:
- Same number of sentences.
- 1% training v2 outperformed 10% training v1 with an eighth of the frames.

Tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Input</th>
<th>Output</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role-prediction</td>
<td>Predicate, arg (head)</td>
<td>role</td>
<td>“child eat apple” → prob of Agent, Patient, ...</td>
</tr>
<tr>
<td>Role/slot-filling (word prediction)</td>
<td>Predicate, role</td>
<td>arg head (lex item)</td>
<td>“child eat Patient” → prob of “apple”, “cake”, ...</td>
</tr>
<tr>
<td>Thematic fit (Padó and McRae norms)</td>
<td>Predicate, arg (head), role</td>
<td>Score [0..1]</td>
<td>“child:Agent eat dog:Patient” → low score for dog in frame+role. Few tests sets; no training data!</td>
</tr>
</tbody>
</table>

Results

<table>
<thead>
<tr>
<th>Size</th>
<th>Ver</th>
<th>Role acc.</th>
<th>Word acc.</th>
<th>final</th>
<th>/Padó max</th>
<th>final</th>
<th>/McRae max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1%</td>
<td>v1</td>
<td>.8857</td>
<td>.0009</td>
<td>.0435</td>
<td>.0001</td>
<td>.2760</td>
<td>.0033</td>
</tr>
<tr>
<td></td>
<td>v2</td>
<td>.9102</td>
<td>.0063</td>
<td>.1029</td>
<td>.0007</td>
<td>.3149</td>
<td>.0038</td>
</tr>
<tr>
<td>1%</td>
<td>v1</td>
<td>.9332</td>
<td>.0006</td>
<td>.0819</td>
<td>.0002</td>
<td>.5150</td>
<td>.0029</td>
</tr>
<tr>
<td></td>
<td>v2</td>
<td>.9656</td>
<td>.0001</td>
<td>.1416</td>
<td>.0002</td>
<td>.4850</td>
<td>.00135</td>
</tr>
<tr>
<td>10%</td>
<td>v1</td>
<td>.9419</td>
<td>.0017</td>
<td>.0941</td>
<td>.0005</td>
<td>.5166</td>
<td>.00345</td>
</tr>
<tr>
<td></td>
<td>v2</td>
<td>.9715</td>
<td>.0010</td>
<td>.1541</td>
<td>.00045</td>
<td>.5229</td>
<td>.00227</td>
</tr>
<tr>
<td>20%</td>
<td>v1</td>
<td>.9445</td>
<td>.0003</td>
<td>.0982</td>
<td>.0011</td>
<td>.5219</td>
<td>.00069</td>
</tr>
<tr>
<td></td>
<td>v2</td>
<td>.9733</td>
<td>.0004</td>
<td>.1621</td>
<td>.00048</td>
<td>.5363</td>
<td>.00305</td>
</tr>
</tbody>
</table>

Research answers
- Training data savings: improving annotation quality reduces data requirement up to 10-fold for role and word prediction.
- Models trained on better lemma identification, better parsing, better SRL tags did better than baseline at all most data sizes.
- We are releasing a large resource with modern annotation: RW-English v2