A Rule-Based Unsupervised Morphology Learning Framework

Constantine Lignos, Erwin Chan*, Mitch Marcus, Charles Yang
University of Pennsylvania, *University of Arizona

Morpho Challenge 2009
CLEF 2009, 9/30/2009
Defining the Task

• Application of a language acquisition model as a morphological analyzer

• How do we define an acquisition model?
  • Cognitively motivated- the representations it learns are linguistically motivated and cognitively useful
  • Designed for a child’s input- Small amounts of sparse data received in an unsupervised fashion

• Not looking to create a fully psychologically plausible algorithm
  • While the structures learned are plausible, some parts of the algorithm are computationally expensive for the sake of simplicity
The Learning Model: Chan (2008)

- **Structures and Distributions in Morphology Learning**
- **Provides:**
  - Representation of morphology- Base and Transforms Model
  - Simple bootstrapping algorithm for learning bases and transforms in an unsupervised fashion
- **Enhancements needed for Morpho Challenge:**
  - Adaptation to larger/noisier corpora
  - Morphological analysis output
  - Support for multi-step derivations
Distribution of Inflected Forms

Log(freq)

Spanish newswire verbs (2.5 M)
Base and Transforms Model

• Within each syntactic category, the most common inflected form is consistent
• Instead of relying on an abstract stem, we have a “base” form that we can easily identify - the most common inflection in each category
• To model a derived form, apply a transform to a base:

\[
\text{RUN} + (\$, s) = \text{runs} \\
\text{MAKE} + (e, \text{ing}) = \text{making}
\]

Note: $ is used to represent a null affix
Base and Transforms Model

- The learner will learn a set of rules (transforms) and the word pairs they apply to (base-derived pairs)

```
Bake
BAKE (Base)

Baker
BAKE + ER ($, er)

Baking
BAKE + ING (e, ing)

Bakes
BAKE + S ($, s)

Bakers
BAKE + ER+ S ($, s)
```
The Algorithm: Sets

- A word belongs to one of three sets at any time:
  - Unmodeled- All words begin in this set
  - Base- Words that are used as a base in a transform and are not derived from anything else
  - Derived- Words that are derived from a base word or another derived word
Core Algorithm

1. Pre-process words and populate the Unmodeled set.
2. Until a stopping condition is met, perform the main learning loop:
   1. Count affixes in words of the (Base + Unmodeled) set and the Unmodeled set.
   2. Hypothesize transforms from words in (Base + Unmodeled) to words in Unmodeled.
   3. Select the best transform.
   4. Reevaluate the words that the selected transform applies to, using the Base, Derived and Unmodeled sets
   5. Move the words used in the transform accordingly.
4. Output analysis
# English Transforms Learned

<table>
<thead>
<tr>
<th>Trans.</th>
<th>Sample Pair</th>
<th>Trans.</th>
<th>Sample Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+($, s)</td>
<td>scream/screams</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>+($, ed)</td>
<td>splash/splashed</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>+($, ing)</td>
<td>bond/bonding</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>+($, ‘s)</td>
<td>office/office’s</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>+($, ly)</td>
<td>unlawful/unlawfully</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>+(e, ing)</td>
<td>supervise/supervising</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>+(y, ies)</td>
<td>fishery/fisheries</td>
<td>21</td>
</tr>
<tr>
<td>8</td>
<td>+($, es)</td>
<td>skirmish/skirmishes</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>+($, er)</td>
<td>truck/trucker</td>
<td>23</td>
</tr>
<tr>
<td>10</td>
<td>($, un)+</td>
<td>popular/unpopular</td>
<td>24</td>
</tr>
<tr>
<td>11</td>
<td>+($, y)</td>
<td>risk/risky</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>($, dis)+</td>
<td>credit/discredit</td>
<td>26</td>
</tr>
<tr>
<td>13</td>
<td>($, in)+</td>
<td>appropriate/ inappropriate</td>
<td>27</td>
</tr>
<tr>
<td>14</td>
<td>+($, ation)</td>
<td>transform/transformation</td>
<td></td>
</tr>
</tbody>
</table>
Performance

English

German

Score (%)

F-Measure
Precision
Recall

Iterations

9/30/2009
Error Types and Proposed Solutions

• Almost all transforms learned are real morphological rules, although they sometimes have spurious pairs
  • In English, +($, a) and ($ ,se)+ are the only spurious transforms out of 27 learned
  • Example spurious pairs for good transforms:
    — gust/disgust
    — pen/penal
    — tent/intent
    — gin/begin
  • Part of the cause is there is no concept of syntactic categories
    — Thus no concept of inflectional/derivational rules
    — Basic approach to category induction in Chan 2008, but needs refinement to identify category of derived forms
Error Types and Proposed Solutions

• **Difficulty learning multistep derivations**
  - Does not predict existence of unseen forms
    - Ex: acidified = ACID + ($, ify) + (y, ied)
    - If *acidify* is not seen in the corpus we won’t learn the connection between *acid* and *acidified*
  - The learner needs to understand the productivity of rules in order to decide whether it’s likely an unseen form exists

• **Rule representation too simple for other languages**
  - All rules consist of affix changes only
  - Should support wider morphological functions, such as templatic morphology and vowel harmony
Conclusions

• An acquisition model can provide an effective learning framework for a morphological analyzer
• Chan (2008) model and algorithm deliver competitive results in English and German with some adaptation
• To cover more languages, the representations used by the learner needs to be expanded