Using Hand-Written Rewrite Rules to Induce Underlying Morphology

Michael A. Tepper

University of Washington
Department of Linguistics

Unsupervised Morpheme Analysis – Morpho Challenge 2007
Outline

Introduction
Morphemes and Allomorphs
Examples from Challenge Languages

Procedure Overview
Rewrite Rules
Stage A :: Basic EM
Stage B :: Split Segments

Results
F-Measure Results

Summary
Definitions

We consider morphemes to be...

▶ basic units of grammar with no internal structure which may be composed together to form words
▶ realized as sequences of linguistic symbols (phones and/or letters)

Morphemes may be rendered differently in different contexts:

▶ lexical context: /s/ → en, as in oxen
▶ phonological/orthographic context: /s/ → es, as in dresses

Morphological variants are known as allomorphs
### Examples

<table>
<thead>
<tr>
<th>Language</th>
<th>Type</th>
<th>Morpheme</th>
<th>Allomorphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>stem</td>
<td>/wake/</td>
<td>wake, wak</td>
</tr>
<tr>
<td></td>
<td>suffix</td>
<td>/s/</td>
<td>s, es</td>
</tr>
<tr>
<td>Finnish</td>
<td>stem</td>
<td>/katto/</td>
<td>katto, kato</td>
</tr>
<tr>
<td></td>
<td>suffix</td>
<td>/ta/</td>
<td>a, ä, ta, tä</td>
</tr>
<tr>
<td>Turkish</td>
<td>stem</td>
<td>/kanad/</td>
<td>kanad, kanat</td>
</tr>
<tr>
<td></td>
<td>suffix</td>
<td>/dik/</td>
<td>dik, dük, dik,</td>
</tr>
</tbody>
</table>
Using Hand-Written Rewrite Rules to Induce Underlying Morphology
Flowchart

Preprocess

Original Wordlist

Morfessor 0.9

Categories-MAP

Propose Underlying Analyses

Estimate HMM Probabilities

Re-segment Wordlist

Rewrite Rules

STAGE A :: EM

STAGE B :: SPLIT

B1

Re-tag Segmentation

B2

Propose Underlying Analyses

B3

Estimate HMM Probabilities

B4

Re-segment (Split) Morphs

Rewrite Rules

surface-layer

analysis-layer

probabilities

Tepper University of Washington

Using Hand-Written Rewrite Rules to Induce Underlying Morphology
Analysis by Rewrite Rules

- Written as cascaded (ordered) rewrite rules and compiled into regular expressions.
- Rules are meant to be run in the analysis direction on a surface segmentation.
- For efficiency, we only permit two types of analyses per segment $s$:
  - analyses where all the rules that could have applied, did. ($u''$)
  - analyses where no rules applied ($u' = s$)
- Example Rule capturing the fact that English suffix /s/ is written as $es$ after sibilants (s, z, sh, ...):

$$\begin{align*}
\varnothing & \quad \rightarrow \quad e \\
\text{underlying} & \quad \quad \text{surface} \\
& \quad \quad / \quad [\text{+SIB}] \quad + \quad _s
\end{align*}$$

(1)
Stage A :: Basic EM

Preprocess
Morfessor 0.9
Categories-MAP

A1 Propose Underlying Analyses

A2 Estimate HMM Probabilities

A3 Re-segment Wordlist

STAGE B :: SPLIT

B1 Re-tag Segmentation

B2 Propose Underlying Analyses

B3 Estimate HMM Probabilities

B4 Re-segment (Split) Morphs

Flowchart

Tepper University of Washington

Using Hand-Written Rewrite Rules to Induce Underlying Morphology
Stage A :: Basic EM

- We estimate transition and emission probabilities of a morfessor-style HMM via maximum likelihood.
- Emission probabilities are estimated by observing cooccurrences of segments $s_i$ in the surface layer, $u_i$ in the analysis layer, with tags $t_i$ to estimate the probability $P(u_i|t_i)$ of emitting underlying morphemes:

$$P(u_i|t_i) = \sum_{s \in \text{allom.-of}(u_i)} P(u_i, s|t_i)$$

Where:

$$u_i = \begin{cases} u'_i & \text{if } u_i = s_i \\ u''_i & \text{otherwise} \end{cases}$$
Stage A :: Basic EM

Find the maximum probability segmentation of the wordlist by finding the argmax of the following equation for each word:

\[
\arg\max_{u,t} P(u|t)P(t) \approx \arg\max_{u,t} \left[ \prod_{i=1}^{n} P(u_i|t_i)P(t_i|t_{i-1}) \right]
\]  (3)
Stage B :: Split Segments

Flowchart

STAGE A :: EM
Preprocess
Morfessor 0.9
Categories-MAP
A1 Propose Underlying Analyses
A2 Estimate HMM Probabilities
A3 Re-segment Wordlist

STAGE B :: SPLIT
B1 Re-tag Segmentation
B2 Propose Underlying Analyses
B3 Estimate HMM Probabilities
B4 Re-segment (Split) Morphs
Stage B :: Split Segments

- **Re-tag** the segmentation first, using Creutz and Lagus’s 2004-2005 heuristic technique, such that only morphs exhibiting prototypical affix- or stem-distributional features are tagged as such.

- The remainder are tagged as **noise**; this makes them unavailable to be used in splitting.

- **Key:** Forcably **split** segments that are too frequent break under normal circumstances.
## F-Measure Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Morf.-CatMAP</td>
<td>82.17%</td>
<td>33.08%</td>
<td>47.17%</td>
</tr>
<tr>
<td></td>
<td>Bernhard2</td>
<td>61.63%</td>
<td>60.01%</td>
<td>60.81%</td>
</tr>
<tr>
<td></td>
<td>Tepper2-b300</td>
<td>75.62%</td>
<td>51.72%</td>
<td>61.43%</td>
</tr>
<tr>
<td></td>
<td><strong>1% impr.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>Morf.-CatMAP</td>
<td>76.83%</td>
<td>27.54%</td>
<td>40.55%</td>
</tr>
<tr>
<td></td>
<td>Bernhard2</td>
<td>59.65%</td>
<td>40.44%</td>
<td>48.20%</td>
</tr>
<tr>
<td></td>
<td>Tepper-b600</td>
<td>62.01%</td>
<td>46.20%</td>
<td>52.95%</td>
</tr>
<tr>
<td></td>
<td><strong>10% impr.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>Zeman</td>
<td>65.81%</td>
<td>18.79%</td>
<td>29.23%</td>
</tr>
<tr>
<td></td>
<td>Morf.-CatMAP</td>
<td>76.36%</td>
<td>24.50%</td>
<td>37.10%</td>
</tr>
<tr>
<td></td>
<td>Tepper-b100</td>
<td>61.15%</td>
<td>49.22%</td>
<td>54.54%</td>
</tr>
<tr>
<td></td>
<td><strong>47% impr.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our approach, which utilizes a small amount of knowledge in an otherwise unsupervised framework, is successful at learning underlying morphology.

Learning improvements over unsupervised approaches are more dramatic for languages with more allomorphic effects, like Turkish (not surprising).

There is hope that with a technique such as ours we can pinpoint generalizations about the most effective rules, which would be useful towards developing features for templates from which to learn rules.
Thank you!

Acknowledgements

Funding
- UW Simpson Center for the Humanities
- UW Graduate School

Thesis Committee
- Dr. Fei Xia
- Dr. Emily Bender

Friends and Colleagues
- Tia Ghose
- Jonathan North Washington

Special Thanks
Morpho Challenge Team
- Dr. Mikko Kurimo
- Dr. Mattias Creutz
- Matti Varjokallio
- Ville Turunen