On the Importance of Subword Information for Morphological Tasks in Truly Low-Resource Languages

Yi Zhu1 Benjamin Heinzerling2,3 Ivan Vulić1 Michael Strube4 Roi Reichart5 Anna Korhonen1
1Language Technology Lab, University of Cambridge 2RIKEN AIP 3Tohoku University 4Heidelberg Institute for Theoretical Studies 5Technion, IIT

tl;dr
► Subwords are great, but how do they perform in low-resource settings?
► This work: thorough analysis of several subword methods
► Three morphological tasks, 16 languages
► Simulated and actual low-resource settings
► Simulate two types of scarcity: scarce embedding training data and scarce task data
► Scarsity of task data has a much larger impact
► No subword method best in all settings, but character n-gram often strongest, followed by BPE.

Subword-informed Word Representations

Subword Methods

Word dishonestly

word dishonestly

dish on est (ly)

chars dishonest

chars dishonest on est (ly)

bpe1e5 dishonest on est (ly)

bpe1e4 dishonest

bpe1e3 dishonest

bpe1e2 dishonest

bpe1e1 dishonest

bpe1e0 dishonest

Three Morphological Tasks

► Fine-grained entity typing (FGET): Lincolnshire → /location/county
Data: Wikidata + Freebase
► Morphological tagging (MTAG): her → (Gen=Fem, Num=Sing, Per=3, Poss=Yes, PronType=Ps)
Data: Universal Dependencies
► Named entity recognition (NER): Barack Obama (→ person) was born in Hawaii (→ location).
Data: WikiAnn

16 Languages: Embedding and Task Data (Tokens)

<table>
<thead>
<tr>
<th>Language</th>
<th>MTAG</th>
<th>NER</th>
<th>FGET</th>
<th>Agglutinative</th>
<th>Fusional</th>
<th>Intro</th>
<th>Isolat</th>
</tr>
</thead>
<tbody>
<tr>
<td>norf</td>
<td>70.84</td>
<td>52.43</td>
<td>52.40</td>
<td>33.74</td>
<td>49.65</td>
<td>49.07</td>
<td>47.09</td>
</tr>
<tr>
<td>charn</td>
<td>77.22</td>
<td>54.20</td>
<td>53.81</td>
<td>34.02</td>
<td>51.32</td>
<td>49.16</td>
<td>47.09</td>
</tr>
<tr>
<td>w2v</td>
<td>70.43</td>
<td>54.20</td>
<td>53.81</td>
<td>34.02</td>
<td>51.32</td>
<td>49.16</td>
<td>47.09</td>
</tr>
<tr>
<td>bert</td>
<td>70.43</td>
<td>54.20</td>
<td>53.81</td>
<td>34.02</td>
<td>51.32</td>
<td>49.16</td>
<td>47.09</td>
</tr>
</tbody>
</table>

Takeaways

► Scarcity of task data has a much larger impact than scarcity of embedding data.
► Subword-informed architectures are better than word-based methods in most cases, particularly in low resource settings (e.g., ZU, BM, GOT).
► When available, multilingual BERT performs well in MTAG and NER, but subword models are better in FGET. Gap becomes smaller or disappears with more embedding training data.
► No one-size-fits-all method, but character n-grams often strongest, followed by BPE.