Cross-lingual tagger evaluation without test data

Željko Agić  
zeag@itu.dk

Barbara Plank  
b.plank@rug.nl

Anders Søgaard  
soegaard@di.ku.dk

Three views on languages

on the web

have annotations

written

with test data

studied at ACL
(thanks, Sebastian Mielke!)

2016

An evaluation impasse

“What we create is affected by what we can evaluate.”
—Winston Churchill 😊

Dictionaries to the rescue!

<table>
<thead>
<tr>
<th>Wiktionary quality</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>86.81</td>
<td>18.38</td>
</tr>
<tr>
<td>$D_{trans}$</td>
<td>58.09</td>
<td>11.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metrics evaluation</th>
<th>acc diff</th>
<th>P@1</th>
<th>$T_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{soft}$</td>
<td>-11.98</td>
<td>14/25</td>
<td>0.42</td>
</tr>
<tr>
<td>$A_{trans}$</td>
<td>-8.85</td>
<td>11/20</td>
<td>0.27</td>
</tr>
</tbody>
</table>