Direct Assessment

Yvette Graham

August 11, 2016
Direct Assessment (DA) I

- Consideration being given to using DA alone for next year

Reasons:

- High correlation between RR and DA
- It seems like we could get good clusters with (conservatively) half the annotation time
- Computed as follows:
  - we require 100 hits per system submission, average 5 min per hit, so 500 minutes = 8 hours
  - DA at 500 translations is what we might need (maybe more in some cases), and that takes about 2.5 hours (half hour per hit)
Direct Assessment (DA) II

- English side can be completely crowdsourced
- Leaves researchers responsible only for tasks where we can’t find crowdsourced workers
Correlation of RR and DA

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>cs-en</td>
<td>0.997</td>
</tr>
<tr>
<td>fi-en</td>
<td>0.996</td>
</tr>
<tr>
<td>tr-en</td>
<td>0.988</td>
</tr>
<tr>
<td>de-en</td>
<td>0.964</td>
</tr>
<tr>
<td>ru-en</td>
<td>0.961</td>
</tr>
<tr>
<td>ro-en</td>
<td>0.920</td>
</tr>
<tr>
<td>en-ru</td>
<td>0.975</td>
</tr>
</tbody>
</table>
# Human Assessment vs Significant Differences: CS-EN
DE-EN

% Significant Differences

Assessed Translations per System

DA

RR (with de-dup.)
RO-EN

% Significant Differences
Assessed Translations per System

DA
RR (with de-dup.)
RU-EN

% Significant Differences

Assessed Translations per System

DA
RR (with de-dup.)
TR-EN

% Significant Differences

DA
RR (with de-dup.)

Assessed Translations per System
Conclusions

- Trial of DA was successful overall
- No problems crowd-sourcing all to-English language pairs
- Not enough workers for all out-of-English news LPs except English to Russian – those LPs unfortunately must remain the task of participants
- Correlation with RR high across the board
- DA almost achieves as many significant differences as RR but without deduplication

More to come:

- WMT’16 included RR with deduplication and DA without it – Makes the comparison of numbers of judgments difficult
- Future: Compare unexpanded (undeduped) versions to see what effect it had, since this is really an unfair comparison