Rob van der Goot
@robvanderg

Abstract
This work explores normalization for parser adaptation. Traditionally, normalization is used as separate pre-processing step. We show that integrating the normalization model into the parsing algorithm is beneficial. To this end, we use a normalization model combined with the parsing as intersection algorithm. This way, multiple normalization candidates can be leveraged, which improves parsing performance on social media. We test this hypothesis by modifying the Berkeley parser; out-of-the-box it reaches an F1 score of 66.52. Our integrated approach performs significantly better, with an F1 score of 67.36, while using the out-of-the-box it reaches an F1 score of 66.94.

Figure 1: The output of the normalization model for the sentence ‘new pix coming tomorrow’.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sent</th>
<th>Words/Unk%</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ (2011)</td>
<td>16,520</td>
<td>15.3 4.4</td>
</tr>
<tr>
<td>Li and Liu (2014)</td>
<td>2,577</td>
<td>15.7 14.1</td>
</tr>
<tr>
<td>SUN (2014)</td>
<td>200</td>
<td>10.9 7.9</td>
</tr>
<tr>
<td>EWT (2-21)</td>
<td>39,832</td>
<td>23.9 4.4</td>
</tr>
<tr>
<td>Foster et al. (2011)</td>
<td>269</td>
<td>11.1 9.3</td>
</tr>
</tbody>
</table>

Table 1: Some basic statistics for our training and development corpora. % of unknown words (Unk) calculated against the Aspell dictionary ignoring capitalization.

Rob van der Goot @robvanderg · Jan 10
The output of the Berkeley parser on a noisy sentence and its automatically normalized counterpart.

*Interesting*

Berkeley parser: 70.85 66.52
Best norm. seq.: 72.04 66.94
Integrated norm.: 72.77 67.36
Gold POS tags: 74.98 71.80

#parseradaptation #naturallanguageprocessing #socialmedia 

Rob van der Goot @robvanderg · Jan 10
That is interesting! maybe we can use the parsing as intersection algorithm to improve even further?

@robvanderg · Jan 15
That is interesting! maybe we can use the parsing as intersection algorithm to improve even further?

Rob van der Goot @robvanderg · Jan 20
Overview of the model:

Rob van der Goot @robvanderg · Jan 22
@GJ F1 scores on the development data when integrating multiple candidates while normalizing ALL words or only the UnKnown words:

Rob van der Goot @robvanderg · Jan 22
@GJ F1 scores on the development data when integrating multiple candidates while normalizing ALL words or only the UnKnown words:

Rob van der Goot @robvanderg · Jan 15
But Rob, is this #Significant?

Rob van der Goot @robvanderg · Jan 10
The output of the Berkeley parser on a noisy sentence and its automatically normalized counterpart.

*Interesting*