A new automatic spelling correction model aimed at improving parsability

Rob van der Goot and Gertjan van Noord
Old approach

- IV/OOV
- Generate candidates
- Rank candidates
New approach

- IV/OOV
- Generate candidates
- Rank candidates
Data

- LexNorm v1.2
- 549 tweets / 10,576 tokens
- 2,140 OOV tokens
- 1,184 tokens corrected
only 3mths left in school.

i will always miss my skull, NO frnds and my teachrs.

new pix comming tomoroe

new pictures coming tomorrow.
IV/OOV

- Aspell dictionary
- IV tokens skipped
- 90% of the errors (Bo Han, 2013)

Example:
- I am tiret
- I am tire
I am tired

I am tired

is amy tired

im aim tire
I am tire

I am tire

is amy tired

im aim tire
Generate candidates

- Edit distances (Modified Aspell)
- N-grams
- Original token
Generate candidates

![Bar chart showing recall against n-gram candidates for 20 ins. Cost and 30 ins. Cost. The x-axis represents n-gram candidates ranging from 100 to 10000, while the y-axis represents recall ranging from 960 to 1140. The chart shows a general increase in recall as the number of n-gram candidates increases.]
Generate candidates
Rank candidates

- N-grams
- Edit distance
- Occurrence in dictionaries
- Parse probability
Rank candidates

1. Random Forest
2. Coordinate Ascent
3. MART
4. RankBoost
5. RankNet
6. AdaRank
7. LambdaMART
8. ListNet
Rank candidates

- Average 222 candidates

<table>
<thead>
<tr>
<th>top</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>5</td>
<td>0.62</td>
</tr>
<tr>
<td>10</td>
<td>0.72</td>
</tr>
</tbody>
</table>
(Dis-) Advantages

- Includes IV errors
- More general
- Adaption

- Less efficient
- Training data
Future work

- Rank on sentence level
- Generate different token orders
- Generate multi-word solutions
- New corpus (parses)