A Large-Scale Query Spelling Correction Corpus
Matthias Hagen, Martin Potthast, Marcel Gohsen, Anja Rathgeber, Benno Stein

Spelling Correction
Crucial part of query understanding pipeline.
Typical errors:
▶ Deletion: entertaner → entertainer
▶ Insertion: baseballl → baseball
▶ Space: sponge bobe → spongebob
▶ Special character: noahs ark → noah’s ark
▶ Substitution: canfederate → confederate
▶ Transposition: chevorlet → chevrolet

Publically available corpora:
▶ Microsoft Speller Challenge 2011
  ▶ 5,892 queries, 19.1% with alternative spelling
  ▶ 811 with potential misspelling (13.8%) 
  ▶ 311 with definite misspelling (5.3%) 
▶ JDB corpus from the qSpell team
  ▶ 6,000 queries, 16.4% with alternative spelling
  ▶ 418 with potential misspelling (7.0%) 
  ▶ 565 with definite misspelling (9.4%) 

Our Corpus
▶ Webis-QSpell-17
  ▶ 54,772 queries, 16.7% with alternative spelling
  ▶ 52,427 with potential misspelling (4.4%) 
  ▶ 56,744 with definite misspelling (12.3%) 
▶ Available at
http://www.uni-weimar.de/medien/webis/corpora/
▶ Construction:
1. 55,555 queries sampled from AOL log (frequencies, lengths, bots)
3. 54,772 queries manually spell-checked by 2 annotators (“tools” allowed)
4. Discussion of disagreements between annotators
5. Queries with alternative spellings double-checked by 3 annotators
6. 9,171 queries with alternative spellings in the end
▶ Remark: Segmentations for almost all queries in companion corpus Webis-QSeC-10

Spelling Checker Evaluation
▶ Spell checkers
  ▶ Baseline: Do nothing
  ▶ Google: Scraped “Did you mean” etc.
  ▶ Bing: Spell Check API
  ▶ Lueck: Reimplementation of Microsoft Speller Challenge winner
▶ Confidence values
  ▶ Spell checkers return confidence for a correction (sum to 1 per query)
▶ Evaluation measures
  ▶ Prec@1: Is the top correction correct?
  ▶ Variants of precision and recall

\[
EP = \frac{1}{|Q|} \sum_{q \in Q} \sum_{c \in C_q} P(c|q)
\]
\[
ER = \frac{1}{|Q|} \sum_{q \in Q} \frac{|C_q \cap G_q|}{|G_q|}
\]
\[
EF_1 = 2 \cdot \frac{EP \cdot ER}{EP + ER}
\]
Q set of queries
C_q set of computed spelling variants for a query q
G_q set of spelling variants in ground truth for a query q
P confidence value of a spelling variant c for a query q
▶ Code available at
https://github.com/webis-de/SIGIR-17

Query spelling correction performance.

<table>
<thead>
<tr>
<th></th>
<th>Prec@1</th>
<th>EF_1</th>
<th>EP</th>
<th>ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft corpus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>0.96</td>
<td>0.89</td>
<td>0.96</td>
<td>0.83</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.95</td>
<td>0.87</td>
<td>0.95</td>
<td>0.81</td>
</tr>
<tr>
<td>Bing</td>
<td>0.95</td>
<td>0.87</td>
<td>0.93</td>
<td>0.81</td>
</tr>
<tr>
<td>Lueck</td>
<td>0.65</td>
<td>0.85</td>
<td>0.89</td>
<td>0.82</td>
</tr>
<tr>
<td>JDB corpus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>0.95</td>
<td>0.91</td>
<td>0.94</td>
<td>0.89</td>
</tr>
<tr>
<td>Bing</td>
<td>0.93</td>
<td>0.89</td>
<td>0.92</td>
<td>0.86</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.91</td>
<td>0.87</td>
<td>0.91</td>
<td>0.84</td>
</tr>
<tr>
<td>Lueck</td>
<td>0.62</td>
<td>0.88</td>
<td>0.90</td>
<td>0.86</td>
</tr>
<tr>
<td>Our corpus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.88</td>
<td>0.86</td>
<td>0.88</td>
<td>0.83</td>
</tr>
<tr>
<td>Bing</td>
<td>0.88</td>
<td>0.84</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>Lueck</td>
<td>0.56</td>
<td>0.85</td>
<td>0.83</td>
<td>0.86</td>
</tr>
</tbody>
</table>

▶ Our corpus seems to be a little harder (Prec@1)
▶ Only Google really outperforms do-nothing baseline
▶ Only Google performs above 0.5 for most error types
▶ Exception: space errors (also Google below 0.5)
▶ Lueck struggles with Prec@1