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 bob` → `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
  - 2,427 with potential misspelling (4.4%)
  - 6,744 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)
  2. Manual removal of non-English and inappropriate queries
  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
- Corpus characteristics (error types with absolute frequencies and percentage of queries with alternative spellings)

	MS	JDB	Ours
Deletion	308 (27.5%)	226 (23.0%)	3,082 (33.6%)
Insertion	163 (14.5%)	235 (23.9%)	1,691 (18.4%)
Space	625 (55.7%)	497 (50.6%)	2,847 (31.0%)
Special character	0 (0.0%)	0 (0.0%)	3,230 (35.2%)
Substitution	135 (12.0%)	118 (12.0%)	1,751 (19.1%)
Transposition	31 (2.8%)	27 (2.8%)	386 (4.2%)

- Remark: Segmentations for almost all queries in companion corpus Webis-QSeC-10

Spell 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

$$\mathrm{EP} = \frac{1}{|Q|} \cdot \sum_{q \in Q} \sum_{c \in C_q \cap G_q} P(c|q)$$

$$\mathrm{ER} = \frac{1}{|Q|} \cdot \sum_{q \in Q} \frac{|C_q \cap G_q|}{|G_q|}$$

$$\mathrm{EF}_1 = 2 \cdot \frac{\mathrm{EP} \cdot \mathrm{ER}}{\mathrm{EP} + \mathrm{ER}}$$

Q	set of queries q
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.

	Prec@1	EF_1	EP	ER
Microsoft corpus				
Google	0.96	0.89	0.96	0.83
Baseline	0.95	0.87	0.95	0.81
Bing	0.95	0.87	0.93	0.81
Lueck	0.65	0.85	0.89	0.82
JDB corpus				
Google	0.95	0.91	0.94	0.89
Bing	0.93	0.89	0.92	0.86
Baseline	0.91	0.87	0.91	0.84
Lueck	0.62	0.88	0.90	0.86
Our corpus				
Google	0.93	0.92	0.93	0.90
Baseline	0.88	0.86	0.88	0.83
Bing	0.88	0.84	0.86	0.83
Lueck	0.56	0.85	0.83	0.86

- 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

Big Data Analytics + Webis
www.matthias-hagen.de + www.webis.de

Bauhaus-Universität Weimar
matthias.hagen@uni-weimar.de