Using Google to Create a More Accurate and Extensible Spelling Corrector

Making a Better Spelling Corrector

Hepburn Best
Hayden Metsky
Colin Sidoti
Jacqueline Somogyi
Jose Rosario
Problems With Existing Spelling Checkers

- Do not use context
- Cannot recognize new words or slang
- Cannot recognize proper nouns
- Difficult to adapt to inflected and agglutinative languages
- Time intensive to adapt to languages with many irregular forms
Intellectual Idea

- Use web search engine results to rank possible spellings
- The web is the world’s largest corpus
- Constantly updated
- Contains words, in context, from almost all written languages
- Large source of proper nouns and slang
Current Spelling Correctors

- Aspell, Ispell, Microsoft Word
  - Scan for Misspellings
  - Suggest Corrections
- Quality determined by Suggestion Intelligence and Suggestion Intelligence First
Example

IE: The Governor's School student *aet* pie for dinner.

<table>
<thead>
<tr>
<th>Suggestion Intelligence</th>
<th>Suggestion Intelligence First</th>
</tr>
</thead>
<tbody>
<tr>
<td>eat</td>
<td>ate</td>
</tr>
<tr>
<td>ate</td>
<td>eat</td>
</tr>
<tr>
<td>tea</td>
<td>tea</td>
</tr>
<tr>
<td>abet</td>
<td>abet</td>
</tr>
<tr>
<td>eta</td>
<td>eta</td>
</tr>
</tbody>
</table>
Traditional Flaws and Our Solution

Commonplace Suggestions
IE: After Governor's School, I will go home and relax.

Three Possible Corrections: wall, well, and will

Google's Result Counts:

“I wall go” 115
“I well go” 7,220
"I will go" 9,230,000
Edit Distance

- **Addition**
  - “gvernor” → “governor”

- **Deletion**
  - “schoool” → “school”

- **Substitution**
  - “engineerinh” → “engineering”

- **Transposition**
  - “tecnhology” → “technology”
The Metrics

- Keyboard Distance
- N-Graphs
- Morphology
- Google Result Counts
Keyboard Distance

A suggestion has a higher probability of being correct if the keyboard distance is lower.

Schook → School = Edit Distance of 1
“K” is 1 key away from “L” on the keyboard.
N-Graphs

- A string of n letters
- We look for the number of occurrences of each n-graph
- Project Gutenberg
- Implemented an n-graph script in Perl
- IE: Governor
- 2-Graphs: go, ov, ve, er, rn, no, or
- 3-Graphs: gov, ove, ver, ern, rne, ner
- Etc...
Morphology

- Builds lists of common prefixes, stems, and suffixes based on an analysis of a corpus
- Uses these to divide a possible correction into prefix + stem + suffix
- Treats each word part as separate from the word as a whole and finds the probability of each part existing alone
- Uses these probabilities of the word parts existing alone to rank possible corrections
Google

- I lost teh robotics competition today

- Google:
  - “the”
  - “tea”

- Pre + Google:
  - “lost the”
  - “lost tea”

- Google + Post:
  - “the robotics”
  - “tea robotics”

- Pre + Google + Post:
  - “lost the robotics”
  - “lost tea robotics”
Final Calculations

- Determine a weight for each metric
- For each possible correction, multiply the weight of the metric by the number the metric returned
- Add all of the above products for each suggestion
- Each possible correction will then result in a probability between 0 and 1
  - The higher the probability, the more likely the suggestion is correct
See It In Action

http://www.movietally.com/hayden/spellchecker/GSET/
Results

- Used a sample set of 47 misspelled words
- 75% of misspelled words were correctly respelled
- 17% of misspelled words were incorrectly respelled
- 8% of misspelled words could not at all be respelled (i.e., no possible corrections could be found)

- Results for Hungarian are pending
Future Work

- Optimization
  - Caching
    - Possible Corrections
    - Google Result Counts
- Quality
  - Frontier Words
    - IE: “I would like a pieceof pie.”
  - Fine Tune Weightings
  - More Edit-Distances
- Usage
  - Adapt to other languages
Acknowledgments

- Blase Ur
- Marc L'Heureux
- Wanda Duran
- Governor’s School- Dean Brown
- Governor’s School Board of Overseers
- Sponsors: Prudential, Morgan Stanley, Rutgers University, The John and Margaret Post Foundation, John and Laura Overdeck
Q+A