Using Learner Data to Improve Error Correction in Adjective-Noun Combinations

Ekaterina Kochmar and Ted Briscoe

The ALTA Institute, Computer Laboratory
University of Cambridge
Error detection and correction (ED&C): State-of-the-art

Attracted much attention recently:

- books [Leacock et al., 2014; Leacock et al., 2010]
- shared tasks [Ng et al., 2014; Ng et al., 2013; Dale et al., 2012, Dale and Kilgarriff, 2011]
- multiple papers and dissertations
- multiple workshops (10th anniversary of BEA!)

However, so far:

- major focus on grammatical errors, errors in articles and prepositions
- fewer address other error types [Kochmar and Briscoe, 2014; Ng et al., 2014; Rozovskaya et al., 2014; Sawai et al., 2013; Dahlmeier and Ng, 2011]
Our work: Focus

Errors in content words (ANs in particular)

- **Frequent error types** [Leacock et al., 2014; Ng et al., 2014]

- cover 20% of learner errors in the CLC [Tetreault and Leacock, 2014]

- notoriously hard to master

- yet, important for successful writing [Leacock and Chodorow, 2003; Johnson, 2000; Santos, 1988]
Content word errors: Challenges

- Lack of strictly defined rules:
  - powerful computer $\leftrightarrow$ strong computer
  - powerful tea $\leftrightarrow$ strong tea

- Sources of confusion:
  - similarity in meaning:
    - powerful $\sim$ strong
  - similarity in spelling:
    - classic $\sim$ classical
  - overusing words with general meaning:
    - big vs broad\|wide\|long\|
  - L1-related confusions
    - good humor vs good mood (cf. French bon humor)
### Function vs Content Words

<table>
<thead>
<tr>
<th>Function words</th>
<th>Content words</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶ Multi-class classification using number of possible alternatives</td>
<td>▶ What are the multiple classes?</td>
</tr>
<tr>
<td>▶ Availability of finite confusion sets</td>
<td>▶ Confusion sets depend on the original word choice</td>
</tr>
<tr>
<td>▶ Error detection and correction – possible to do simultaneously</td>
<td>▶ Error detection independent of error correction [Kochmar and Briscoe, 2014]</td>
</tr>
</tbody>
</table>
Basic EDC algorithm

Three-step algorithm [Leacock et al., 2014]:

1. \( \forall X \) look for more fluent/native-like \( Y \)'s
2. compare \( Y \)'s to \( X \) using some frequency-based measure
3. if \( \exists Y_i \) more fluent than \( X \) \( \Rightarrow \) \( X \) is an error, \( Y_i \) is a correction
Basic EDC algorithm performance

- The choice of the metric for ranking
  - Quality of the system suggestions

- The choice of the source of alternatives
  - Coverage of the EC algorithm
Different sources

- Reference databases of known learner errors and their corrections [Wible et al., 2003; Shei and Pain, 2000]
- Semantically related: WordNet, dictionaries and thesauri [Östling and Knutsson, 2009; Futagi et al., 2008; Shei and Pain, 2000]
- Spelling alternatives and homophones [Dahlmeier and Ng, 2011]
- L1-specific confusion sets [Dahlmeier and Ng, 2011; Chang et al., 2008; Liu, 2002]
- Wikipedia revisions [Madnani and Cahill, 2014]
In this work
We treat error detection and error correction as separate steps, and focus on error correction

Contributions

1. Explore different ways to construct the correction sets and to rank the alternatives
2. Demonstrate how error patterns extracted from learner text can be used to improve the ranking of the alternatives
3. Present an EDC system for AN combinations
4. Explore the usefulness of augmenting sets of alternatives for an EC system
Datasets

1. the AN dataset extracted from the Cambridge Learner Corpus (CLC) and annotated with respect to the learner errors
   http://ilexir.co.uk/media/an-dataset.xml

2. the AN dataset extracted from the CLC-FCE dataset
   http://ilexir.co.uk/applications/adjective-noun-dataset/

3. the AN dataset extracted from the CoNLL-2014 Shared Task on Grammatical Error Correction training and development sets
   http://www.comp.nus.edu.sg/~nlp/conll14st.html
## Datasets

### Annotated dataset
- **340** unique errors
- annotated with the error types for adjectives and nouns (S, F and N)

### CLC-FCE dataset
- **456** ANs that have adjective–noun combinations as corrections
- no annotation for error types

### NUCLE dataset
- **369** ANs that have adjective–noun combinations as corrections
- no annotation for error types
- smaller number of L1s, different set of topics, etc.
### Distribution of errors in the choice of adjectives (A), nouns (N) or both words

<table>
<thead>
<tr>
<th>Word</th>
<th><strong>Ann. data</strong></th>
<th><strong>CLC-FCE</strong></th>
<th><strong>NUCLE</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>63.24%</td>
<td>43.20%</td>
<td>34.15%</td>
</tr>
<tr>
<td>N</td>
<td>30.29%</td>
<td>52.63%</td>
<td>60.16%</td>
</tr>
<tr>
<td>Both</td>
<td>6.47%</td>
<td>4.17%</td>
<td>5.69%</td>
</tr>
</tbody>
</table>

---

Cambridge ALTA

Ekaterina Kochmar and Ted Briscoe

Error Correction in Adjective–Noun Combinations
Key points

1. Explore resources to retrieve alternatives and report \textit{coverage}.
   \begin{itemize}
   \item \textit{coverage} – proportion of gold standard corrections covered by the resources
   \end{itemize}

2. Rank AN alternatives and assess the \textit{quality} of ranking (\textit{MRR}).
   \begin{itemize}
   \item \textit{quality} – ability of the algorithm to rank the more appropriate corrections higher than the less appropriate ones
   \end{itemize}

3. Use confusion sets extracted from the learner data.
Levenshtein distance (Lv): form-related confusions, F
E.g.: *electric society → electronic society
important *costumer → important customer

WordNet (WN): semantically related confusions, S
E.g.: *heavy decline → steep decline
good *fate → good luck

Confusion pairs from the CLC: cover L1-related confusions, N
E.g.: *strong noise → loud noise
historical *roman → historical novel
### Coverage

**Coverage of different sets of alternatives**

<table>
<thead>
<tr>
<th>Setting</th>
<th>Ann. data</th>
<th>CLC-FCE</th>
<th>NUCLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lv</td>
<td>0.1588</td>
<td>0.0833</td>
<td>0.0897</td>
</tr>
<tr>
<td>WN</td>
<td>0.4353</td>
<td>0.3904</td>
<td>0.2880</td>
</tr>
<tr>
<td>CLC</td>
<td>0.7912</td>
<td>0.8684</td>
<td>0.5625</td>
</tr>
<tr>
<td>CLC+Lv</td>
<td>0.7971</td>
<td>0.8706</td>
<td>0.5951</td>
</tr>
<tr>
<td>CLC+WN</td>
<td>0.8558</td>
<td>0.8904</td>
<td>0.6141</td>
</tr>
<tr>
<td>All</td>
<td><strong>0.8618</strong></td>
<td><strong>0.8925</strong></td>
<td><strong>0.6467</strong></td>
</tr>
</tbody>
</table>
Alternative ANs

\[
\{\text{alternative ANs}\} = \bigcup (\{\text{alternative adjs}\} \times \text{noun}) \\
\cup (\text{adj} \times \{\text{alternative nouns}\})
\]

Evaluation

\[
MRR = \frac{1}{|N|} \sum_{i=1}^{N} \frac{1}{\text{rank}_i}
\]

\(N\) – total number of erroneous ANs
**Ranking measures**

1. *Frequency in the BNC+ukWaC*

2. *Normalised pointwise mutual information (NPMI):*

\[ NPMI(AN) = \frac{PMI(AN)}{-\log_2(P(AN))} \]  

where

\[ PMI(AN) = \frac{\log P(AN)}{P(A)P(N)} \]
## Quality (I)

### MRR for the alternatives ranking (I)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Ann. set</th>
<th>CLC-FCE</th>
<th>NUCLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{CLC}_{freq}$</td>
<td>0.3806</td>
<td>0.3121</td>
<td>0.2275</td>
</tr>
<tr>
<td>$\text{CLC}_{NPMI}$</td>
<td>0.3752</td>
<td>0.2904</td>
<td>0.1961</td>
</tr>
<tr>
<td>$(\text{CLC}+\text{Lv})_{freq}$</td>
<td>0.3686</td>
<td>0.3146</td>
<td>0.2510</td>
</tr>
<tr>
<td>$(\text{CLC}+\text{Lv})_{NPMI}$</td>
<td>0.3409</td>
<td>0.2695</td>
<td>0.1977</td>
</tr>
<tr>
<td>$(\text{CLC}+\text{WN})_{freq}$</td>
<td>0.3500</td>
<td>0.2873</td>
<td>0.2267</td>
</tr>
<tr>
<td>$(\text{CLC}+\text{WN})_{NPMI}$</td>
<td>0.3286</td>
<td>0.2552</td>
<td>0.1908</td>
</tr>
<tr>
<td>$\text{All}_{freq}$</td>
<td>0.3441</td>
<td>0.2881</td>
<td>0.2468</td>
</tr>
<tr>
<td>$\text{All}_{NPMI}$</td>
<td>0.3032</td>
<td>0.2407</td>
<td>0.1943</td>
</tr>
</tbody>
</table>
Exploitation of confusion probabilities

Use the **confusion probabilities** (CP) from the CLC – probabilities associated with the words used as corrections given the original (incorrect) word choice.

**Formula refinement**

\[
M' = M \times CP(a_{\text{orig}} \rightarrow a_{\text{alt}}) \\
\times CP(n_{\text{orig}} \rightarrow n_{\text{alt}}) \tag{5}
\]

- \(M\) – a measure of choice
- \(CP(a_{\text{orig}} \rightarrow a_{\text{alt}=\text{orig}})\) and \(CP(n_{\text{orig}} \rightarrow n_{\text{alt}=\text{orig}})\) set to 1.0
Example: *big enjoyment → great pleasure

<table>
<thead>
<tr>
<th>Original</th>
<th>Alternatives</th>
<th>CP(\text{orig} → \text{alt})</th>
</tr>
</thead>
<tbody>
<tr>
<td>*big</td>
<td>great</td>
<td>0.0144</td>
</tr>
<tr>
<td></td>
<td>large</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>wide</td>
<td>0.0043</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>5.1122 \times 10^{-5}</td>
</tr>
<tr>
<td>enjoyment</td>
<td>pleasure</td>
<td>0.0938</td>
</tr>
<tr>
<td></td>
<td>entertainment</td>
<td>0.0313</td>
</tr>
<tr>
<td></td>
<td>fun</td>
<td>0.0104</td>
</tr>
<tr>
<td></td>
<td>happiness</td>
<td>0.0052</td>
</tr>
</tbody>
</table>
Example: *big enjoyment $\rightarrow$ great pleasure

Basic ranking algorithm (raw frequency)

**System:** great fun (7759 in the native corpus)

**GS:** great pleasure (2829 in the native corpus)

Refined ranking algorithm (frequency’)

**System** & **GC:** great pleasure ($Freq' = 3.8212$)

great fun ($Freq' = 1.1620$)

Freq’ vs freq

fluency in the native data $+$ appropriateness of a correction
### Quality (II)

#### MRR for the alternatives ranking (II)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Ann. set</th>
<th>CLC-FCE</th>
<th>NUCLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{CLC}_{freq}$</td>
<td>0.3806</td>
<td>0.3121</td>
<td>0.2275</td>
</tr>
<tr>
<td>$\text{CLC}_{NPMI}$</td>
<td>0.3752</td>
<td>0.2904</td>
<td>0.1961</td>
</tr>
<tr>
<td>$(\text{CLC}+\text{Lv})_{freq}$</td>
<td>0.3686</td>
<td><strong>0.3146</strong></td>
<td><strong>0.2510</strong></td>
</tr>
<tr>
<td>$(\text{CLC}+\text{Lv})_{NPMI}$</td>
<td>0.3409</td>
<td>0.2695</td>
<td>0.1977</td>
</tr>
<tr>
<td>$(\text{CLC}+\text{WN})_{freq}$</td>
<td>0.3500</td>
<td>0.2873</td>
<td>0.2267</td>
</tr>
<tr>
<td>$(\text{CLC}+\text{WN})_{NPMI}$</td>
<td>0.3286</td>
<td>0.2552</td>
<td>0.1908</td>
</tr>
<tr>
<td>All$_{freq}$</td>
<td>0.3441</td>
<td>0.2881</td>
<td>0.2468</td>
</tr>
<tr>
<td>All$_{NPMI}$</td>
<td>0.3032</td>
<td>0.2407</td>
<td>0.1943</td>
</tr>
<tr>
<td>All$_{freq}'$</td>
<td><strong>0.5061</strong></td>
<td><strong>0.4509</strong></td>
<td><strong>0.2913</strong></td>
</tr>
<tr>
<td>All$_{NPMI}'$</td>
<td>0.4843</td>
<td>0.4316</td>
<td>0.2118</td>
</tr>
</tbody>
</table>
Further analysis of the results

1. Breakdown of the results
   - Top N coverage
   - Error types

2. System augmentation

3. Error detection + correction
% of errors covered by top $N$ alternatives

<table>
<thead>
<tr>
<th>Top $N$</th>
<th>Ann. data</th>
<th>CLC-FCE</th>
<th>NUCLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41.18</td>
<td>34.21</td>
<td>21.20</td>
</tr>
<tr>
<td>2</td>
<td>49.12</td>
<td>45.18</td>
<td>27.99</td>
</tr>
<tr>
<td>3</td>
<td>56.77</td>
<td>50.88</td>
<td>33.70</td>
</tr>
<tr>
<td>4</td>
<td>61.77</td>
<td>55.04</td>
<td>38.04</td>
</tr>
<tr>
<td>5</td>
<td>65.29</td>
<td>58.55</td>
<td>40.49</td>
</tr>
<tr>
<td>6</td>
<td>66.18</td>
<td>61.40</td>
<td>42.39</td>
</tr>
<tr>
<td>7</td>
<td>67.35</td>
<td>62.28</td>
<td>43.21</td>
</tr>
<tr>
<td>8</td>
<td>68.53</td>
<td>63.60</td>
<td>44.29</td>
</tr>
<tr>
<td>9</td>
<td>69.71</td>
<td>65.35</td>
<td>45.38</td>
</tr>
<tr>
<td>10</td>
<td>71.18</td>
<td>66.45</td>
<td>46.20</td>
</tr>
<tr>
<td>Not found</td>
<td>25.29</td>
<td>19.96</td>
<td>48.64</td>
</tr>
</tbody>
</table>
Error type analysis for the annotated dataset

<table>
<thead>
<tr>
<th>Type</th>
<th>S</th>
<th>F</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MRR_{found}$</td>
<td>0.6007</td>
<td>0.8486</td>
<td>0.6507</td>
</tr>
<tr>
<td>Not found</td>
<td>0.1990</td>
<td>0.1705</td>
<td>0.5410</td>
</tr>
</tbody>
</table>

Some observations

- type N (non-related confusion) – the hardest to correct (not surprisingly...)
- type F (form-related) – the easiest (smaller confusion sets)
  e.g., $MRR = 0.875$ for the ANs with *elder*:
  
  elder $\rightarrow$ elderly or older
**NUCLE results**

- 35% of the GS corrections not covered by any sets of alternatives
- confusion sets from the CLC can only cover about 56%
- more limited number of L1s
- different set of topics and learner levels
- more of the type N?

*architectural development → infrastructural development*

*medical *debt → medical bill*  

(6)
Augmenting sets of alternatives

Method

- Add *bill* to the set of alternatives for *debt*
- Add *infrastructural* to the set of alternatives for *architectural*
- ...
- Check whether the results of the error correction system improve

Augmented sets of alternatives results

<table>
<thead>
<tr>
<th>Setting</th>
<th>Ann. set</th>
<th>CLC-FCE</th>
<th>NUCLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLC</td>
<td>0.3806</td>
<td>0.3121</td>
<td>0.2275</td>
</tr>
<tr>
<td>CLC+Lv</td>
<td>0.3686</td>
<td>0.3146</td>
<td>0.2510</td>
</tr>
<tr>
<td>Augm</td>
<td>0.4420</td>
<td>0.3533</td>
<td>0.2614</td>
</tr>
</tbody>
</table>
Algorithm

- **Error detection** ([Kochmar and Briscoe, 2014]):
  \[ P = 0.6850, \quad R = 0.5849 \]
  on the incorrect examples in the annotated dataset

- **Error correction** step:
  - \( MRR = 0.2532 \) on the set of detected errors
  - 24.28% cases GS correction not found
  - \( MRR_{\text{found}} = 0.6831 \)
In this work we:
- focused on EC in adjective–noun combinations
- experimented with 3 publicly available datasets
- looked at the coverage of resources and the quality of suggestions

and we showed:
1. the confusion patterns from the learner data provide the highest coverage and improve the overall ranking
2. error correction system can reach an $MRR$ of 0.5061
3. correction set augmentation is helpful
4. $MRR$ of 0.2532 on the set of errors identified by ED algorithm
Contact: Ekaterina.Kochmar@cl.cam.ac.uk

Data:
- annotated AN dataset
  http://ilexir.co.uk/media/an-dataset.xml
- the AN dataset extracted from the CLC-FCE
  http://ilexir.co.uk/applications/adjective-noun-dataset/
References


T. Santos, 1988. Professors' reaction to the academic writing of nonnative speaking students. TESOL Quarterly, 22(1)


