Not normalized  Need norm.

TN  FP  FN  TP
A Taxonomy for In-depth Evaluation of Normalization for User Generated Content

Rob van der Goot, Rik van Noord & Gertjan van Noord
University of Groningen
r.van.der.goot@rug.nl

26-01-2018
Lexical Normalization

- Pre-tokenized
- Word-word replacements
Lexical Normalization

RT <USERNAME> I mis bein antisocial :
RT <USERNAME> I miss being antisocial :

nt havin friens was the bestest
not having friends was the best
Taxonomy

Why?
- Find strengths/weaknesses normalization models
- Test effect different categories of normalization actions on other task
- Filter out undesirable categories
Taxonomy

Got it?

RT <USERNAME> I mis bein antisocial :

RT <USERNAME> I miss being antisocial :

X X

nt havin friens was the bestest
not having friends was the best

X X X X X
Taxonomy

Unintentional

Word-word

Typographical

Missing

Spelling apostrophe

Split

Merge

Phrasal abbreviation
Taxonomy

Anomalies

Intentional

Unk

Phrasal abbreviation
Repetition
Shortening
Transformations
Slang

Vowels
End
Other
Phonetical
Regular

Other

End

Vowels

Phrasal abbreviation
Repetition
Shortening
Transformations
Slang

Vowels
End
Other
Phonetical
Regular

Taxonomy

Anomalies
- Unintentional
  - Word-word
    - Typographical
    - Missing apostrophe
  - Split
  - Merge
  - Phrasal abbreviation
- Intentional
  - Repetition
  - Shortening
- Transformations
  - Vowels
  - End
  - Other
- Slang
  - Phonetical
  - Regular
- Other
- End
- Vowels

Unk
Taxonomy

1. Typographical error
   spirite $\mapsto$ spirit, complaing $\mapsto$ complaining, throwg $\mapsto$ throw

2. Missing apostrophe
   im $\mapsto$ i’m, yall $\mapsto$ y’all, microsofts $\mapsto$ microsoft’s

3. Spelling error
   favourite $\mapsto$ favorite, dieing $\mapsto$ dying, theirselves $\mapsto$ themselves

4. Split
   pre order $\mapsto$ preorder, screen shot $\mapsto$ screenshot

5. Merge
   alot $\mapsto$ a lot, nomore $\mapsto$ no more, appstore $\mapsto$ app store

6. Phrasal abbreviation
   lol $\mapsto$ laughing out loud, pmsl $\mapsto$ pissing myself laughing

7. Repetition
   soooo $\mapsto$ so, weiiiiird $\mapsto$ weird
8. Shortening vowels
   pls→please, wrked→worked, rmx→remix
9. Shortening end
   gon→gonna, congrats→congratulations, gi→girl
10. Shortening other
    cause→because, smth→something, tl→timeline,
11. Phonetic transformation
    hackd→hacked, gentille→gentle, rizky→risky
12. Regular transformation
    foolin→fooling, wateva→whatever, droppin→dropping
13. Slang
    cuz→because, fina→going to, plz→please
14. Unknown
    skepta→sunglasses, putos→photos
Dataset:

- LexNorm2015
- Shared Task 2nd Workshop on Noisy user-generated Text (W-NUT)
- Train: 2,950 Tweets / 44,385 words
- 8.9% normalized
Annotation

Guidelines

- On unique replacement pairs
- First annotator: all train data (1,204 pairs)
- Second annotator: 150 replacement pairs
- One category per pair
Disagreements: ($\kappa = 0.807$)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>7</td>
<td>difffff</td>
<td>different</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>t</td>
<td>to</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>talkn</td>
<td>talking</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>custa</td>
<td>custand</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>shat</td>
<td>shit</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>tunee</td>
<td>tune</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>downgrdng</td>
<td>downgrading</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>thx</td>
<td>thank</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>yur</td>
<td>your</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>yor</td>
<td>your</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>wearr</td>
<td>wear</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>gf</td>
<td>girlfriend</td>
<td></td>
</tr>
</tbody>
</table>
Most social people are troublesome .. lol

most social people are troublesome .. laughing out loud

0 0 short vow phon. 0 .. phrasal abbrev.
Annotation

% of tokens

Category

Typo
Missing apo
Spelling err
Split
Merge
Phrasal abbr.
Repetition
Short. Vow
Short end
Short other
Phon.
Regular
Slang
Unk

0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

17 / 23
Evaluation MoNoise

kheb da gzien

tokenize

orig → kheb
lookup → ik heb dat gezien
w2v → hb kga
aspell → heb khee

+ N-grams

classifier

ik heb dat gezien
Evaluation MoNoise

www.let.rug.nl/rob/monoise
Evaluation MoNoise

Correctly normalized
Not correct

Category
Typo
Missing apo
Spelling err
Split
Merge
Phrasal abbr.
Repetition
Short, Vow
Short end
Short other
Phon.
Regular
Slang
Unk

number of tokens
Evaluation MoNoise

The diagram illustrates the evaluation of MoNoise, categorizing errors as correctly normalized and not correct. The categories include Typo, Missing apo, Spelling err, Split, Merge, Phrasal abbr., Repetition, Short, Vow, Short end, Short other, Phon., Regular, Slang, and Unk. The y-axis represents the number of tokens, and the bars indicate the distribution of correctly normalized and not correct tokens across different categories.
Evaluation MoNoise

https://bitbucket.org/robvanderg/normtax
https://bitbucket.org/robvanderg/monoise
Extensions

- Effect of different normalization modules on different categories
- Classify other corpora automatically?
- Test effect of each category on end-task (automatic vs gold)