

A Taxonomy for In-depth Evaluation of Normalization for User Generated Content

Lexical Normalization

social ppl r anoying
social people are annoying

new pix comming tomorro
new pictures coming tomorrow

idk proolly like wednesday
i don't know probably like wednesday

Annotation

orig	I teering up from alletgies , cant tell why LMAO			
norm	i tearing up from allergies , can't tell why laughing my ass off			
ann1	Typo	Typo	Missin'	Phrasal abbr.
ann2	Spelling	Typo	Missin'	Phrasal abbr.

$$\kappa = 0.807$$

Dataset: LexNorm2015 (Baldwin et al. 2015)
Available at: www.bitbucket.org/robvander/normtax

Motivation

- Examine strengths and weaknesses of lexical normalization models
- Test the effect of different categories on end-task
- Train a normalization model handling only desired categories

MoNoise

TRY IT:

www.let.rug.nl/rob/monoise



Unknown

putos→photos
skepta→sunglasses

Unintentional



Intentional



Typo

spirite→spirit
complaing→complaining

Missing'

im→i'm
yall→y'all

Spelling

dieing→dying
theirselves→themselves

Split

pre order→preorder
screen shot→screenshot

Merge

alot→a lot
nomore→no more

Phrasal abbr.

lol→laughing out loud
pmsl→pissing myself laughing

Repetition

soooo→so
weiiiiird→weird

Shortening vowel

pls→please
rmx→remix

Shortening end

g→girl
gon→gonna

Shortening other

cause→because
smth→something

Phonetic transf.

hackd→hacked
rizky→risky

Regular transf.

foolin→fooling
droppin→dropping

Slang

fina→going to
cuzi→because

Performance

