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

Rob van der Goot, Rik van Noord & Gertjan van Noord

Lexical Normalization

social ppl r anoying
social people are annoying
new pix comming tomorrow
new pictures coming tomorrow
idk prolly like wednesday
i don’t know probably like wednesday

Annotation

\[ \kappa = 0.807 \]

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

Social ppl r anoying

Performance

TRY IT:
www.let.rug.nl/rob/monoise

TRY IT:
www.let.rug.nl/rob/monoise