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

rijksuniversiteit groningen

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

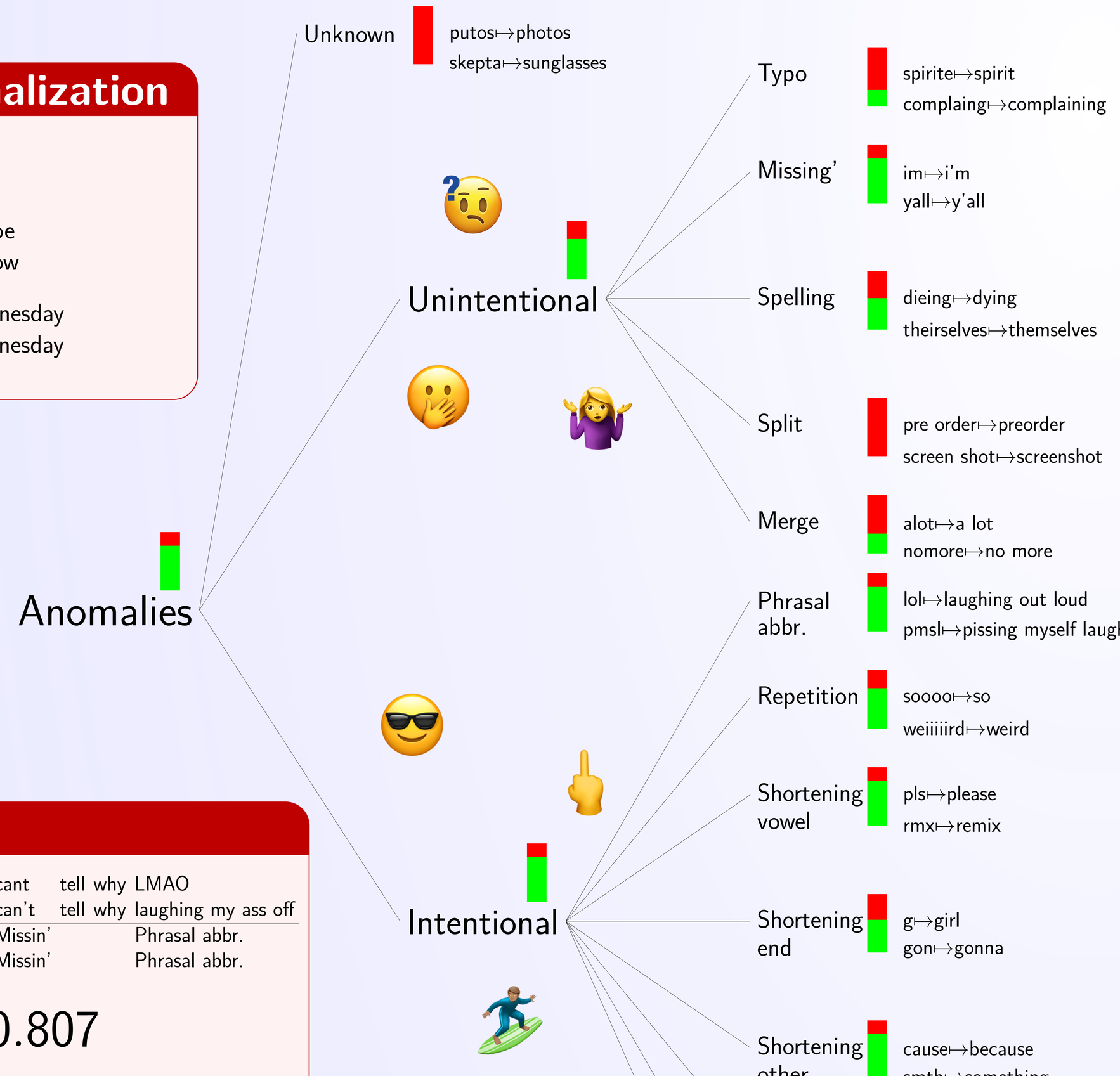
NUANCE FOUNDATION

Lexical Normalization

social ppl r anoying
social people are annoying

new pix comming tomorroe
new pictures coming tomorrow

idk prolly 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.

$$\kappa = 0.807$$

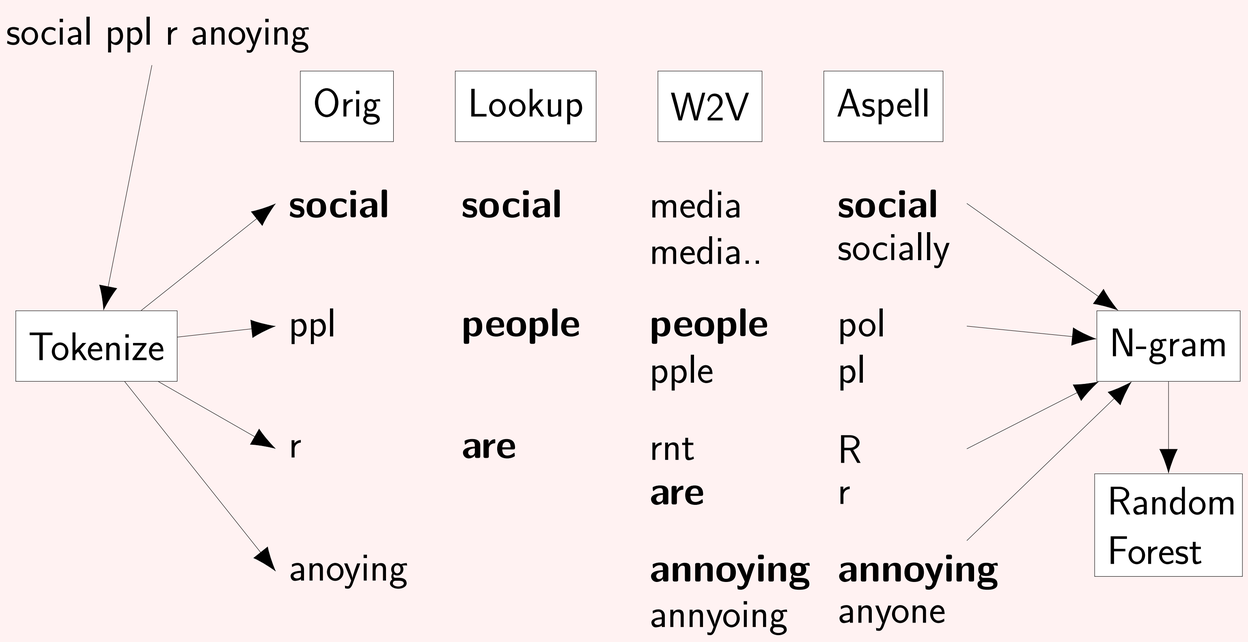
Dataset: LexNorm2015 (Baldwin et al. 2015)
Available at: `www.bitbucket.org/robvanderg/normtax`

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

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



Performance