Lexical Normalization for Dutch Social Media Texts

rijksuniversiteit groningen

Rob van der Goot & Gertjan van Noord

NUANCE FOUNDATION

Lexical Normalization

nee ! :-D kzal nog es vriendenlijk doen lol
nee ! :-D ik zal nog eens vriendelijk doen lol

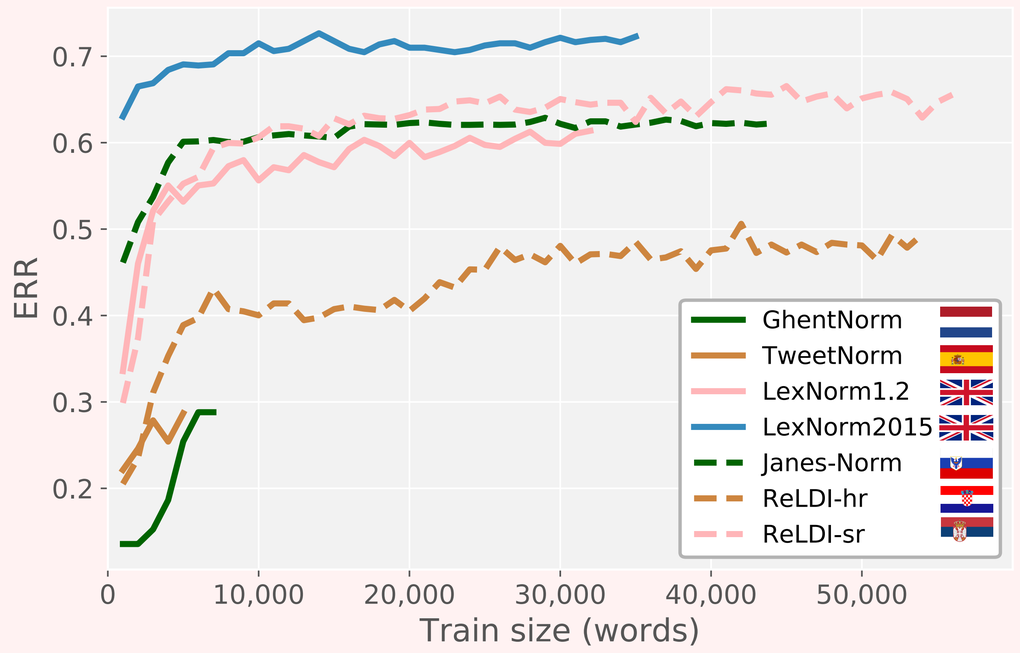
tgaat goed , vdg rustig aaan .
Het gaat goed , vandaag rustig aan .

social ppl r anoying
social people are annoying

aaah buenoo esqe digo pa qe madrugara este jajaja
ah bueno es que digo para qué madrugará este jajaja

nekomu je sarkazm detektor crknu
nekomu je sarkazem detektor crknil

Performance Per Corpus



Is normalization for Dutch more difficult?

MoNoise

Generation

Ranking

Tokenizer

Orig. Word	
scoren	#fail
scoren	#fail

LookupList	
alst	hahahaha
als het	haha

word2vec	
mss	dinnetje
mssn	dinnie
missch	vriendinnetje
misschien	dinnetjes

Aspell	
grapjee	felicteren
grapje	feliciteren
grapjes	flecteren
greepje	fluctueren

Split	
kheb	datis
k heb	dat is

word.*	
ech	waarsch
echt	waarschijnlijk
echo	waarschuw
echeltje	waarschuwt

Features:	
isOrig	N-grams Wiki
Word2vec dist.	N-grams Twitter
Aspell dist.	dict
isSplit	length
word.*	containsAlpha
	origFeats

Random Forest Classifier

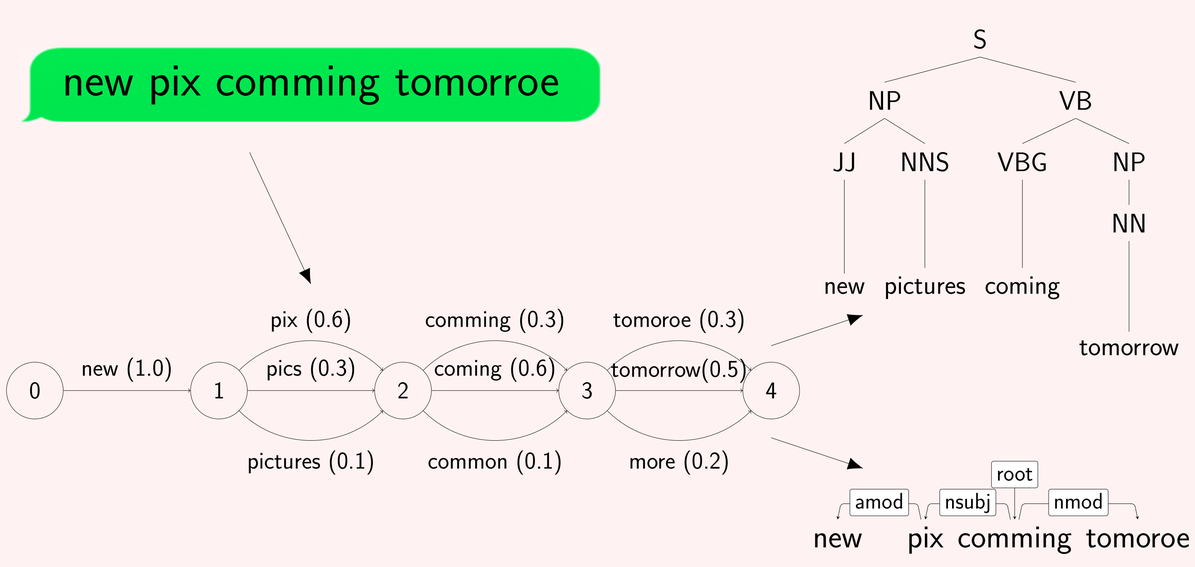
Rob van der Goot and Gertjan van Noord. *MoNoise: Modeling Noise Using a Modular Normalization System.* In CLIN Journal 2017

New Dataset

- Annotate capitalization consistently
- Annotate tokenization in a separate layer
- Do not include phrasal abbreviations
  ('lol'↦'laughing out loud')
- Make publicly available
- No Flemish ↦ Dutch
- Annotate POS dev/test data
- Annotate categories?[a]
- Annotate Universal Dependencies?

[a] Rob Van der Goot, Rik Van Noord, and Gertjan Van Noord. *A taxonomy for in-depth evaluation of normalization for user generated content.* In Proceedings of LREC 2018

Beneficial for Parsing?



Rob Van der Goot and Gertjan Van Noord. *Parser Adaptation for Social Media by Integrating Normalization.* In Proceedings of ACL 2017

Rob Van der Goot and Gertjan Van Noord. *Modeling Input Uncertainty in A Neural Network Dependency Parser.* In Proceedings of EMNLP 2018 **Brussels**

Try it!

www.let.rug.nl/rob/monoise