MoNoise: A Multi-lingual and Easy-to-use Lexical Normalization Tool.

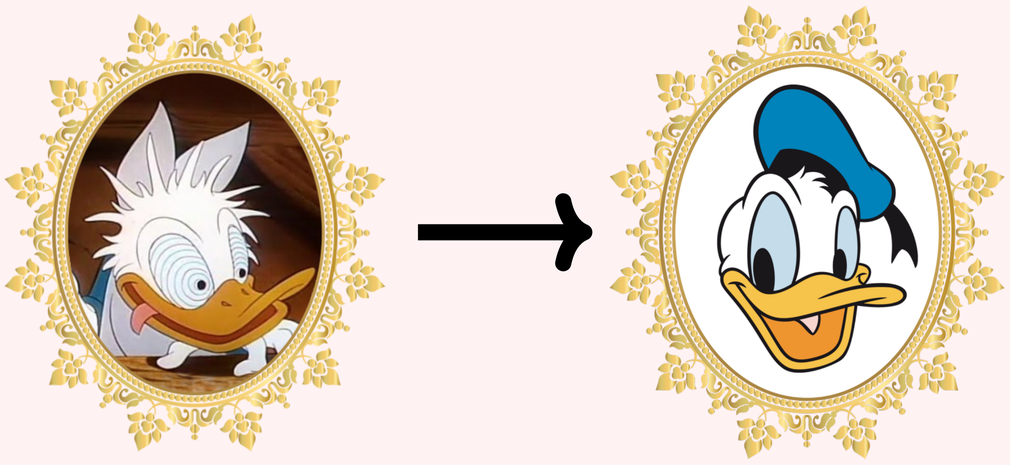
Rob van der Goot

university of groningen

NUANCE FOUNDATION

www.robvandergoot.com/monoise

Lexical Normalization



most social pple r troublesome
most social people are troublesome

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

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

nekomu je sarkazm detektor crknu
nekomu je sarkazem detektor crknil

O simdi usuyordur isitmak lazim
O şimdi üşüyordur ısıtmak lazım

MoNoise

Novel benchmark

Corpus	Words	Lang.	%norm	1-N	Caps	Source
GhentNorm	12,901	NL	4.8	+	+-	[3]
TweetNorm	13,542	ES	6.3	+	+-	[1]
LexNorm1.2	10,576	EN	11.6	-	-	[14]
LiLiu	40,560	EN	10.5	-	+-	[7]
LexNorm2015	73,806	EN	9.1	+	-	[2]
IWT	38,918	TR	8.5	+	+	[5]
Janes-Norm	75,276	SL	15.0	-	+-	[4]
ReLDI-hr	89,052	HR	9.0	-	+-	[9]
ReLDI-sr	91,738	SR	8.0	-	+-	[10]

Error Reduction Rate:

$$ERR = \frac{Accuracy_{system} - Accuracy_{baseline}}{1.0 - Accuracy_{baseline}} \quad (1)$$

- Easily interpretable: percentage of problem solved
- Cross-corpus comparison
- Evaluates complete normalization task

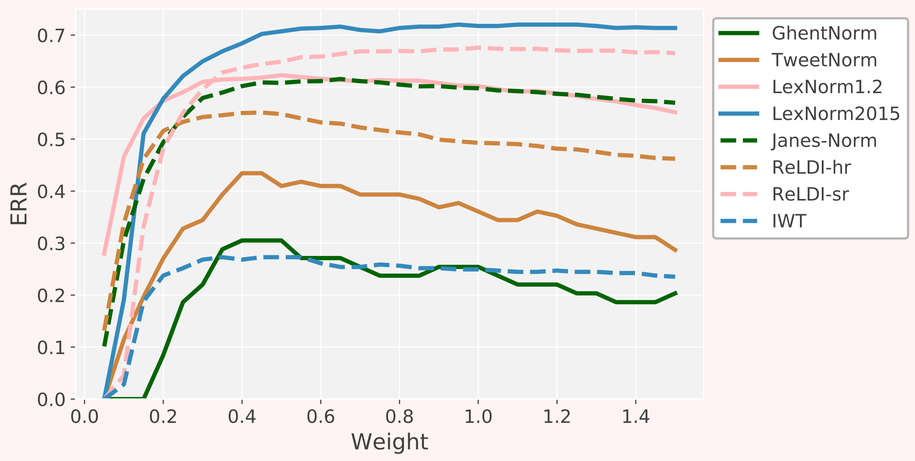
Evaluation

Corpus	Lang	ERR	Prev.	Metric	Prev.	MoNoise
GhentNorm	NL	44.62	[13]	WER	3.2	1.36*
TweetNorm	ES	38.73	[12]	OOV-Prec.	63.4	70.40
LexNorm1.2	EN	59.21	[8]	OOV Acc.	87.58	87.63
LexNorm2015	EN	77.09	[6]	F1	84.21	86.58
IWT	TR	28.94	[5]	OOV Acc.	67.37	48.99
Janes-Norm	SL	31.67	[11] L1	CER	0.38	0.53
Janes-Norm	SL	63.90	[11] L3	CER	1.58	2.24
ReLDI-hr	HR	51.65				
ReLDI-sr	SR	64.61				

* not directly comparable

Novelties

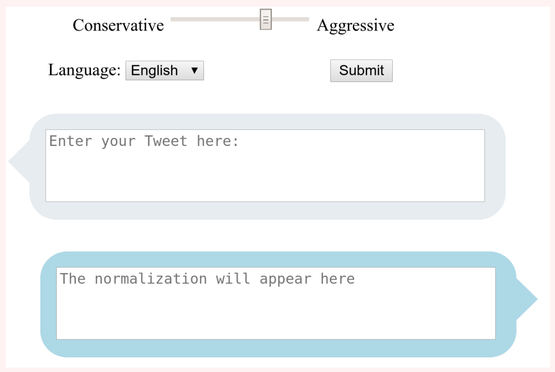
- Much lower ram (En: 18gb↦3gb)
- Better handling of capitalization (if its annotated)
- Features from original word are re-used
- Cached embeddings[a]
- Tune aggressiveness:



[a] see also: https://bitbucket.org/robvanderg/cacheembeds/

Interfaces

Online



www.robvandergoot.com/monoise

Command line

```
--badspeller      --mode=<arg>
--cands=<arg>     --nThreads=<arg>
--caps            --nTrain=<arg>
--dir=<arg>       --output=<arg>
--dev=<arg>       --rf=<arg>
--errdet=<arg>    --seed=<arg>
--feats=<arg>     --syntactic
--feats2=<arg>    --tokenize
--gold            --trees=<arg>
--Gold            --unk=<arg>
--help            --verbose
--input=<arg>     --weight=<arg>
--known=<arg>     --wordline
--kfold=<arg>
```

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

Future languages

Bibliography

[1] Inaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluis Padró, Inaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. Introducción a la tarea compartida Tweet-Norm 2013: Normalización léxica de tuits en español. In *Tweet-Norm@ SEPLN*, 2013.

[2] Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of WNUT*, 2015.

[3] Orphée De Clercq, Sarah Schulz, Bart Desmet, and Véronique Hoste. Towards shared datasets for normalization research. In *Proceedings of LREC*, 2014.

[4] Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Katja Zupan. CMC training corpus janes-tag 2.0, 2017.

[5] Gülşen Eryiğit, Torunoğ-Selamet, and Dilara. Social media text normalization for turkish. *Natural Language Engineering*, 23(6):835–875, 2017.

[6] Ning Jin. NCSU-SAS-Ning: Candidate generation and feature engineering for supervised lexical normalization. In *Proceedings of WNUT*, 2015.

[7] Chen Li and Yang Liu. Improving text normalization via unsupervised model and discriminative reranking. In *Proceedings of the ACL 2014 Student Research Workshop*, 2014.

[8] Chen Li and Yang Liu. Joint POS tagging and text normalization for informal text. In *Proceedings of the Twenty-Fourth IJCAI*, 2015.

[9] Nikola Ljubešić, Tomaž Erjavec, Maja Miličević, and Tanja Samardžić. Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0, 2017.

[10] Nikola Ljubešić, Tomaž Erjavec, Maja Miličević, and Tanja Samardžić. Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0, 2017.

[11] Nikola Ljubešic, Katja Zupan, Darja Fišer, and Tomaz Erjavec. Normalising slovene data: historical texts vs. user-generated content. *Bochumer Linguistische Arbeitsberichte*, 2016.

[12] Jordi Porta and José-Luis Sancho. Word normalization in Twitter using finite-state transducers. *Tweet-Norm@ SEPLN*, 2013.

[13] Sarah Schulz, Guy De Pauw, Orphée De Clercq, Bart Desmet, Véronique Hoste, Walter Daelemans, and Lieve Macken. Multimodular text normalization of Dutch user-generated content. *ACM Transactions on Intelligent Systems Technology*, 2016.

[14] Yi Yang and Jacob Eisenstein. A log-linear model for unsupervised text normalization. In *Proceedings of EMNLP*, 2013.