Lexical Normalization for Code-switched Data and its Effect on POS Tagging

Rob van der Goot and Özlem Çetinoğlu

Lexical normalization

social ppl r troublesome

Lexical normalization

social	ppl	r	troublesome
social	people	are	troublesome

Code-switched data

ak . luv u :(till die

@username tamam yinede gute nacht .

Code-switched data

ak	.	luv	u	:(	till	die
aku	.	love	you	:(	till	die

@username	tamam	yinede	gute	nacht	.
@username	Tamam	yine de	gute	Nacht	.

Datasets

	#words	%norm	% split	%merge	CMI
Id-En	18,758	14.13	1.33	0.17	28.20
Tr-De	13,217	25.97	3.01	1.04	22.44

Table: Descriptive normalization and code-switching statistics on the training split of the datasets.

Datasets

Turkish-German:

- Based on data from Çetinoğlu and Çöltekin (2016)
- Provide alligned normalization
- Also allign LangID's and POS tags

Previous work

Lexical normalization:

- Rule-based: Barik et al. (2019)
- "Traditional" approaches: MoNoise (van der Goot, 2019), Jin (2015)
- Neural networks: Adouane et al. (2019); Lourentzou et al. (2019); Bhat et al. (2018)
- BERT: Muller et al. (2019)

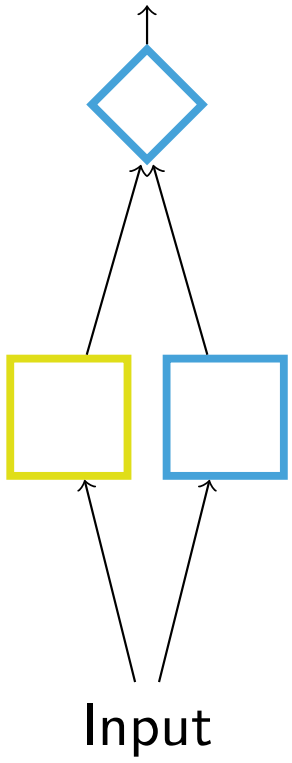
Previous work

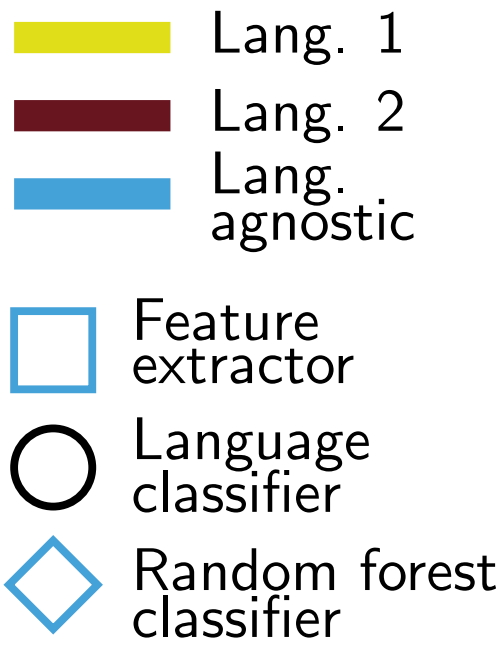
We use MoNoise:

- SOTA for many languages
- Especially powerful in low-resource scenario's
- Open-source

Previous work

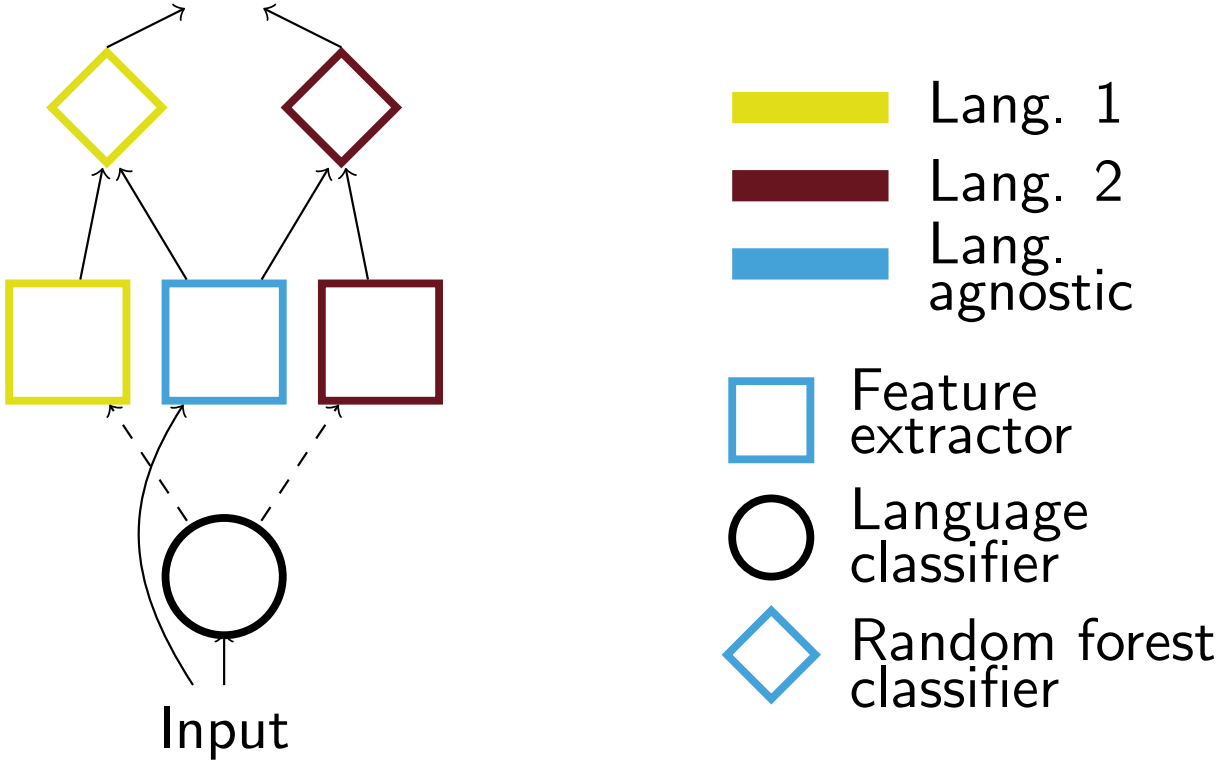
MoNoise:





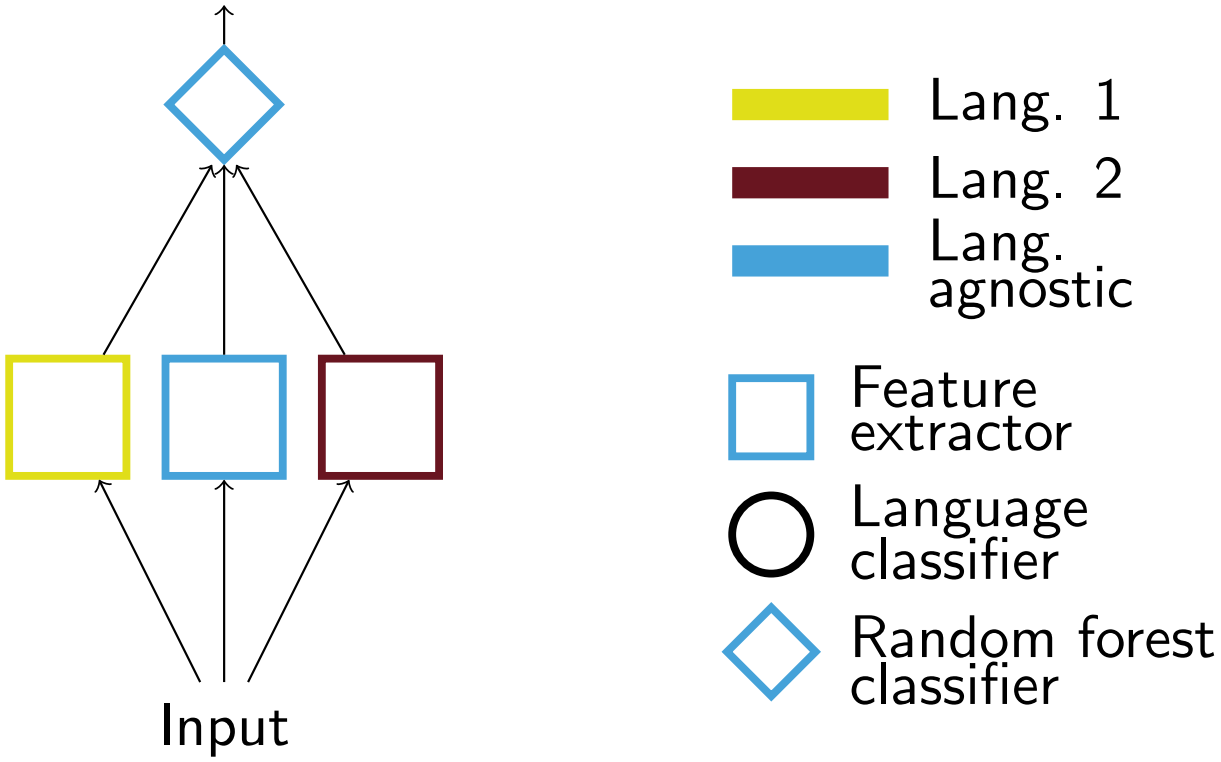
Models

Fragment-based MoNoise:



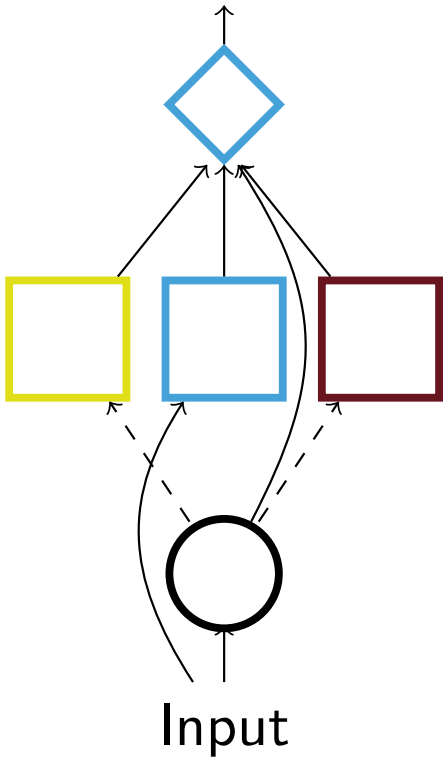
Models

Multilingual MoNoise:



Models

Language-aware MoNoise:



Results

Baselines:

- Leave-As-Is (LAI): return input
- Most-Frequent-Replacement (MFR): return the most frequent normalization for each word learned from training data

Results

Metric: accuracy

Results

Model	Id-En	Tr-De
LAI	73.24	74.03
MFR	*88.35	*78.57
Monolingual-lang1 (Tr/Id)	*94.76	*79.81
Monolingual-lang2 (De/En)	*94.31	80.58
Fragments	94.73	*81.24
Multilingual	94.84	*81.74
Language-aware	94.79	81.68

Table: Normalization performance of the baselines and the proposed models (10-fold accuracy). * Statistical significant.

Results

POS tagging:

- MaChAmp (van der Goot et al., 2021), trained on UD_German-GSD and UD_Turkish-IMST
- Comparison of no normalization, all our models and gold normalization

Results

Model	Tr-De
LAI	60.77
Monolingual (Id/De)	*63.47
Multilingual	*64.06
Language-aware	*63.92
Gold	*67.75

Table: POS tagging accuracies.

Results

Analysis:

- More increase in precision
- LangID performance matters
- More gains on German/English
- Qualitative analysis
- POS confusion matrices

Results

Thanks!

- Code: https://bitbucket.org/robvanderg/csmonoise
- Data: https://github.com/ozlemcek/TrDeNormData

W. Adouane, J.-P. Bernardy, and S. Dobnik. Normalising non-standardised orthography in Algerian code-switched user-generated data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 131–140, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5518. URL https://www.aclweb.org/anthology/D19-5518.

A. M. Barik, R. Mahendra, and M. Adriani. Normalization of Indonesian-English code-mixed twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5554. URL https://www.aclweb.org/anthology/D19-5554.

I. Bhat, R. A. Bhat, M. Shrivastava, and D. Sharma. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1090. URL https://www.aclweb.org/anthology/N18-1090.

Ö. Çetinoğlu and Ç. Çöltekin. Part of speech annotation of a Turkish-German code-switching corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 120–130, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1714. URL https://www.aclweb.org/anthology/W16-1714.

N. Jin. NCSU-SAS-Ning: Candidate generation and feature engineering for supervised lexical normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 87–92, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W15-4313.

I. Lourentzou, K. Manghnani, and C. Zhai. Adapting sequence to sequence models for text normalization in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 335–345, 2019. URL https://ojs.aaai.org//index.php/ICWSM/article/view/3234/3102.

B. Muller, B. Sagot, and D. Seddah. Enhancing BERT for lexical normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 297–306, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5539. URL https://www.aclweb.org/anthology/D19-5539.

R. van der Goot. MoNoise: A multi-lingual and easy-to-use lexical normalization tool. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 201–206, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3032. URL https://www.aclweb.org/anthology/P19-3032.

R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, and B. Plank. Massive Choice, Ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *2005.14672v3*, 2021. URL https://arxiv.org/abs/2005.14672v3.