Addendum for CL-MoNoise: Cross-lingual Lexical Normalization

Rob van der Goot
IT University
robv@itu.dk

1 Modification

MoNoise depends on multiple modules for generation of candidates. One important and language-specific module is the lookup list, which simply remembers all normalization candidates for a given word found in the training data. This module is most probably less effective in cross-lingual setups, so it would make sense to disable it. In the original paper, this was not done, even though it is relatively easy to implement.[1] I have added this in the source code of the original paper.[2], and provide the results in this addendum.

2 Results

Table 1 shows the results of both models on the dev data for all language pairs. It can be seen that only for languages with very close datasets (HR, SL, TR), the version with the lookup list (top) performs better, which is because the lookup lists for these language pairs would have some overlap

Results on the test data are shown in Table **??**, this confirms that CL-MoNoise performs much better without the lookup list.

[1] Simply adding `-f 110111111111` to the training command.

[2] https://bitbucket.org/robvanderg/cl-monoise/src/master/

Target	Source											
	DA	DE	EN	ES	HR	ID-EN	IT	NL	SL	SR	TR	TR-DE
DA	–	2.30	-0.99	-0.59	6.51	-39.71	3.09	-6.38	**8.35**	3.02	-5.06	-0.99
DE	1.01	–	2.44	1.67	0.66	-9.27	10.36	-0.97	-3.96	6.91	12.14	**15.32**
EN	-39.31	1.20	–	-9.34	-45.57	-11.97	**2.74**	-39.38	-41.75	-27.91	-61.85	-50.60
ES	-33.27	-2.53	4.16	–	-32.73	-172.30	**5.06**	-48.82	-67.99	-54.25	-104.90	-19.89
HR	-3.45	-7.03	-3.50	-31.51	–	-106.90	-2.35	-40.34	4.20	**28.54**	-43.25	-36.75
ID-EN	-9.20	5.60	6.91	5.69	-8.47	–	0.54	5.79	-15.82	1.36	**11.68**	7.40
IT	-10.26	-1.62	**2.05**	-5.40	-27.86	-51.84	–	-19.01	-17.93	-20.73	-37.90	-8.53
NL	6.43	17.78	11.03	2.53	6.59	-2.80	15.17	–	8.25	12.64	**19.42**	17.02
SL	1.54	0.90	2.47	-16.90	5.53	-55.32	1.78	-7.93	–	**6.65**	-15.06	-8.71
SR	-2.72	3.26	-4.35	-30.27	**14.77**	-87.43	-0.38	-58.34	-5.88	–	-32.92	-15.81
TR	0.67	9.64	1.22	0.25	4.82	-0.80	5.37	6.25	-2.10	6.00	–	**21.43**
TR-DE	-1.30	8.24	1.91	1.04	-1.88	-2.04	4.05	2.95	-7.01	1.04	**16.54**	–
	Without lookup list											
	DA	DE	EN	ES	HR	ID-EN	IT	NL	SL	SR	TR	TR-DE
DA	–	7.76	1.71	-4.27	8.55	-1.97	3.09	-8.35	**14.27**	12.49	-0.07	3.02
DE	4.38	–	5.97	-12.45	-0.27	-0.78	3.34	-7.25	-0.12	3.92	**28.32**	14.43
EN	1.35	-14.03	–	12.53	-46.06	-8.03	3.45	-40.40	-47.37	-30.16	**23.26**	19.77
ES	-2.35	-11.03	-2.53	–	-33.45	-121.30	0.36	-64.56	-45.03	-60.94	-11.39	**7.96**
HR	0.00	-6.08	-3.81	0.00	–	-53.17	0.00	-60.73	10.73	**26.33**	4.68	0.62
ID-EN	0.00	6.13	4.96	2.29	-7.98	–	0.00	7.88	-10.75	-2.53	**14.21**	4.33
IT	2.92	0.43	2.70	-38.34	-28.73	-17.06	–	-20.30	-16.31	-16.85	**25.59**	9.18
NL	8.52	12.01	9.70	-1.03	7.05	1.17	2.29	–	11.49	12.34	**23.47**	15.06
SL	0.27	-0.56	2.97	0.27	**9.74**	-16.17	0.00	-20.77	–	6.11	0.00	0.54
SR	0.40	-2.20	-4.21	-0.02	**24.67**	-50.07	0.00	-80.69	0.28	–	15.48	1.87
TR	0.96	1.43	2.10	-11.87	5.07	-0.04	2.98	4.86	0.55	7.00	–	**10.94**
TR-DE	0.94	1.26	2.08	0.16	-3.60	-1.36	5.90	0.06	-4.12	1.82	**27.73**	–

Table 1: Cross-lingual performance of MoNoise on training splits (ERR).

Model	Avg.	DA	DE	EN	ES	HR	ID-EN	IT	NL	SL	SR	TR	TR-DE
MoNoise	49.02	51.27	46.96	74.35	45.53	52.63	59.79	21.78	49.53	61.91	59.58	28.21	36.72
MFR	38.37	49.68	32.09	64.93	25.57	36.52	61.17	16.83	37.70	56.71	42.62	14.53	22.09
CL-MoNoise-noLookup	20.99	18.35	30.41	22.33	10.60	25.77	13.06	38.61	20.68	10.57	27.80	7.43	26.29
CL-MoNoise	12.05	7.28	16.55	4.13	4.99	26.41	2.41	0.00	16.22	8.77	20.09	17.57	20.16
LAI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MaChAmp	-21.25	-88.92	-93.36	50.99	25.36	42.62	39.52	-312.87	1.49	56.80	39.44	-12.67	-3.42

Table 2: Results of the models provided by the organizers (grey) and our proposed models. The source language used for CL-MoNoise is shown in the last row.