

A Appendix

A.1 Evaluation

This section provides performance tables that did not fit in the main submission. Tables begin on the next page.

Language	BPE	25000	50000	75000	100000	125000	150000	175000	200000
Czech	16000	18.70	23.90	25.80	27.40	28.70	29.40	31.00	30.70
Czech	32000	17.50	23.50	26.50	28.50	29.40	30.20	31.20	31.50
German	16000	20.00	24.80	26.70	28.90	28.80	29.50	30.60	31.20
German	32000	18.50	23.90	27.00	28.40	29.80	30.00	31.00	32.00

Table 1: MT test set lowercase BLEU by language, joint BPE merge operations, and number of training sentences.

Lang.	Collection	Model	BPE	25000	50000	75000	100000	125000	150000	175000	200000
Czech	Europarl	BM25	16000	0.316	0.395	0.416	0.433	0.457	0.453	0.467	0.477
Czech	Europarl	BM25	32000	0.306	0.406	0.444	0.460	0.476	0.482	0.491	0.498
Czech	Wikipedia	BM25	16000	0.218	0.303	0.327	0.340	0.356	0.359	0.375	0.374
Czech	Wikipedia	BM25	32000	0.141	0.247	0.294	0.314	0.333	0.338	0.350	0.349
Czech	Wikipedia	Neural	16000	0.061	0.084	0.092	0.098	0.104	0.112	0.116	0.114
Czech	Wikipedia	Neural	32000	0.042	0.071	0.086	0.093	0.099	0.101	0.107	0.112
German	Europarl	BM25	16000	0.337	0.416	0.438	0.470	0.461	0.481	0.483	0.500
German	Europarl	BM25	32000	0.319	0.414	0.453	0.475	0.494	0.503	0.500	0.518
German	Wikipedia	BM25	16000	0.301	0.365	0.384	0.396	0.400	0.408	0.413	0.416
German	Wikipedia	BM25	32000	0.185	0.310	0.354	0.374	0.390	0.392	0.397	0.408
German	Wikipedia	Neural	16000	0.080	0.103	0.113	0.119	0.115	0.121	0.125	0.121
German	Wikipedia	Neural	32000	0.056	0.082	0.103	0.108	0.111	0.115	0.118	0.123

Table 2: IR test set RBO by language, collection, IR model, joint BPE merge operations, and number of training sentences.

Lang.	Collection	Model	BPE	25000	50000	75000	100000	125000	150000	175000	200000
Czech	Europarl	BM25	16000	0.254	0.337	0.357	0.376	0.391	0.395	0.405	0.411
Czech	Europarl	BM25	32000	0.240	0.332	0.377	0.389	0.404	0.407	0.420	0.424
Czech	Wikipedia	BM25	16000	0.434	0.585	0.616	0.642	0.666	0.675	0.699	0.692
Czech	Wikipedia	BM25	32000	0.278	0.498	0.573	0.613	0.634	0.642	0.661	0.661
Czech	Wikipedia	Neural	16000	0.045	0.063	0.072	0.083	0.086	0.091	0.090	0.091
Czech	Wikipedia	Neural	32000	0.028	0.056	0.069	0.069	0.078	0.079	0.083	0.082
German	Europarl	BM25	16000	0.273	0.350	0.379	0.399	0.396	0.413	0.419	0.430
German	Europarl	BM25	32000	0.250	0.352	0.396	0.412	0.422	0.435	0.433	0.444
German	Wikipedia	BM25	16000	0.606	0.702	0.732	0.752	0.753	0.774	0.775	0.779
German	Wikipedia	BM25	32000	0.380	0.614	0.689	0.717	0.742	0.747	0.758	0.772
German	Wikipedia	Neural	16000	0.068	0.091	0.101	0.105	0.094	0.106	0.107	0.108
German	Wikipedia	Neural	32000	0.038	0.062	0.090	0.096	0.098	0.095	0.105	0.107

Table 3: IR test set MAP by language, collection, IR model, joint BPE merge operations, and number of training sentences.