Supplementary Material for the Paper: "How to Probe Sentence Embeddings in Low-Resource Languages: On Structural Design Choices for Probing Task Evaluation"

Anonymous EMNLP submission

A Quality of Machine Translated Data

	BLEU (1-gram)	METEOR	MoverScore
en-ka	0.271	0.149	0.272
en-ru	0.470	0.335	0.353
en-tr	0.493	0.359	0.398
en-de	0.446	0.342	0.337

Table 1: Quality of the machine translation service used to translate training data for downstream tasks on reference datasets.

We automatically translated the input data for the AM and TREC downstream tasks. To estimate the quality of the machine translated data, we measured the performance of the service used to translate the data with the help of the JW300 corpus (Agić and Vulić, 2019; Tiedemann, 2012). For each of the language pairs `en-ka`, `en-tr`, and `en-ru`, we translated the first 10,000 sentences of the respective bitext files from JW300 and measured their quality in terms of BLEU, METEOR and MoverScore (Zhao et al., 2019).[1] As reference, we also added `en-de` as a pair with two well-resourced languages. Results are summarized in Table 1. They show that, with the exception of `en-ka`, all language pairs have high-quality translations (on par or even better than `en-de`). We thus expect the influence of errors of the machine translated data to be minimal in `tr` and `ru`. For `ka`, this is not necessarily the case.

B Sentence Encoder Dimensions

Table 2 shows the full list of encoders used in our study and their dimensionalities.

C Class Imbalance

In addition to the classifier type and size, we also tested the influence of the class (im)balance of the training data. In particular, for the four binary probing tasks BigramShift, SubjNumber, SV-Agree, and Voice, we examine the effect of imbalancing with ratios of 1:5 and 1:10. We use LR with sizes of 10k, 20k, and 30k training instances and correlate

[1] Misaligned sentences were skipped.

Encoder	Size
Avg	300
pmeans (Avg+Max+Min)	900
Random LSTM	4096
InferSent	4096
QuickThought	2400
LASER	1024
BERT	768

Table 2: Encoders and their dimensionalities.

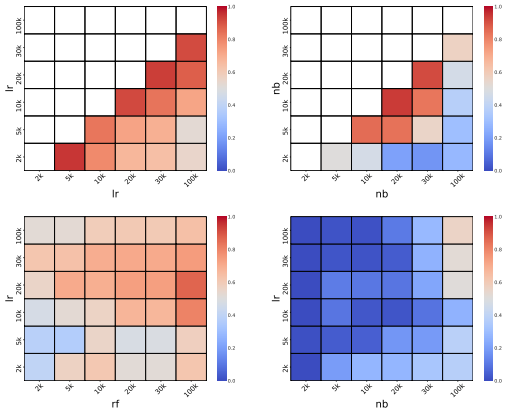

Figure 1: Top: Average correlations $\text{sim}_c(s,t)$ for LR (left) and NB (right), using Spearman. Bottom: Average correlations $\text{sim}_{c,d}(s,t)$ for c = LR and d = RF (left) and c = LR and d = NB (right).

	LR	RF	NB
MLP	.531/.777	.312/.564	-.071/.205
LR		.362/.602	-.107/.147
RF			-.111/.287

Table 3: Min/Avg values $\text{sim}_{c,d}(s,t)$ across (s,t) (using Spearman) between classifiers c and d.

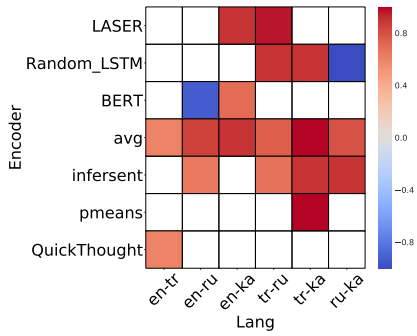

Figure 2: Spearman correlations across languages for different encoders.

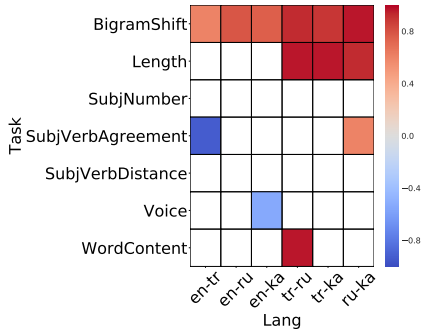

Figure 3: Spearman correlations across languages for different probing tasks.

the results for imbalanced datasets with the standardly balanced datasets. We find that (i) for two tasks (BigramShift, SV-Agree) there is typically high correlation (0.6-0.8) while for the other two tasks the correlation is typically zero between the balanced and imbalanced setting; (ii) correlation to the setting 1:1 (slightly) diminishes as we increase the class imbalance from 1:5 to 1:10. Thus, the scenarios 1:5 and 1:10 do not strongly correlate with 1:1 (as used in all our other experiments). As a consequence, in the multilingual setup, we paid attention to keep datasets as uniform as possible.

D Spearman correlations

Figures 1–4 and Table 3 show Spearman correlation results for the corresponding Pearson results in the main text.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

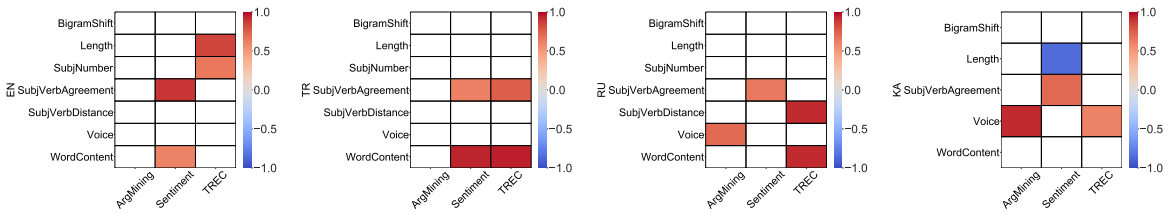

Figure 4: Spearman correlation among probing task and downstream performance for all languages.