Due to lack of monolingual non-English evaluation data, we consider the transfer from other languages to English (lang. → en).

- replace the source language vocabulary with the English vocabulary.
- map each vocabulary token of the target language to the trained source embedding in the same position by order of frequency.
- continue pre-training the model on English Wikipedia.
- fine-tune on tasks from the GLUE benchmark.

- All transferred models improve over the randomly initialised English BERT model, independently of the source language.
- The pre-training data size in the source language has a positive effect on the model performance in the target language.

### Method

- Our baseline is a BERT-base model, namely [en], that is trained from scratch on our English corpus using the same hyperparameters with the transferred models.

### References