# Cross-lingual Transfer of Monolingual Models

Evangelia Gogoulou[1], Ariel Ekgren[2], Tim Isbister[2], Magnus Sahlgren [2]

[1]RISE, [2]AI Sweden

{evangelia.gogoulou}@ri.se

{ariel.ekgren, tim.isbister, magnus.sahlgren}@ai.se

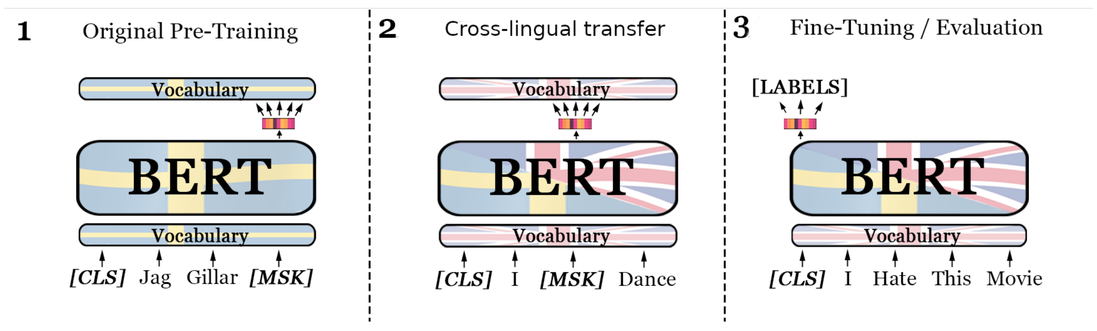
## Introduction

- Training a language model from scratch requires considerable amount of data and computational resources.
- Previous work has mostly focused on cross-lingual transfer from multilingual models [5, 4, 9, 2].
- Recent studies in cross-lingual learning [7, 3, 1] have cast doubt on the previous hypothesis that shared vocabulary and joint pre-training are necessary conditions for cross-lingual generalization.

- Which is the downstream effect of adapting existing monolingual models to a target language, in comparison with a model trained from scratch on the same language?
- Does the effectiveness of crosslingual transfer depend on the similarity between source and target language?

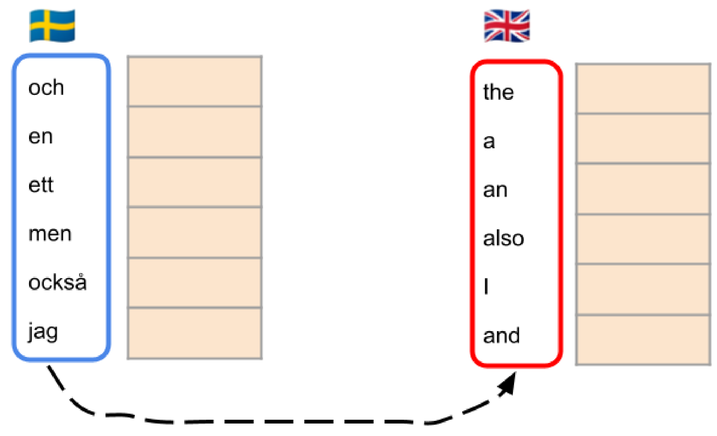
## Method

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




For each language pair, we take a pre-trained language model in a source language and:

- *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




- 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.

| Language | Model name | Alias | Vocab size | Data (GB) |
|---|---|---|---|---|
| English | BERT-base (ours) | en | 32,000 | 13 |
| Swedish | KB-BERT | sv | 50,325 | 18 |
| Dutch | BERTje | nl | 30,000 | 12 |
| Finnish | FinBERT | fi | 50,105 | ≈ 48 |
| Arabic | AraBERTv01 | ar1 | 64,000 | 23 |
| | AraBERTv02 | ar2 | 64,000 | 77 |

Table 1: List of the monolingual BERT models considered. Data size refers to the size of data used for pre-training.

## GLUE Results

| Lang | CoLA | MNLI (m/mm) | MRPC | QNLI | QQP | RTE | SST-2 | STS-B | AVG |
|---|---|---|---|---|---|---|---|---|---|
| en | 25.68 | 76.21/76.13 | 82.69 | 85.86 | 85.21 | 54.21 | 88.04 | 82.71 | 72.97 |
| sv→ en | 43.65 | 80.72/81.77 | 88.93 | 89.11 | 86.32 | 55.08 | 90.22 | 84.91 | **77.85** |
| nl→ en | 39.87 | 78.96/79.79 | 85.65 | 87.34 | 85.82 | 55.01 | 89.10 | 83.65 | **76.13** |
| fi→ en | 40.01 | 79.90/80.52 | 87.82 | 88.30 | 86.37 | 52.12 | 88.18 | 83.82 | **76.34** |
| ar1→ en | 33.29 | 78.90/79.38 | 87.16 | 87.46 | 86.09 | 54.21 | 88.87 | 84.46 | **75.54** |
| ar2→ en | 39.82 | 79.52/80.28 | 88.46 | 88.35 | 85.72 | 57.18 | 90.10 | 83.77 | **77.02** |

Table 2: Average GLUE validation score for all models and tasks excluding WNLI.

- 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.

## English Probing

- **Syntactic**: We evaluate the word representations yielded by [lang→ en] on the *structural probe* model, proposed by [6], which detects whether syntactic trees are encoded in a linear transformation of the model embedding space. The English part of Universal Dependencies v2.7 is used for training.
- **Semantic**: Probing is done using the Words in Context (WiC) task [8], where the model needs to determine if a given word is used with the same meaning or not in two different contexts.

| | *Syntax* | | *Semantics* |
|---|---|---|---|
| **Language** | **UUAS** | **DSpr** | **WiC** (acc) |
| en | 66.21 | 70.41 | 56.73 |
| sv→ en | 67.22 | 72.15 | 61.09 |
| nl→ en | 66.59 | 71.73 | 59.46 |
| fi→ en | 67.02 | 71.56 | 61.06 |
| ar1→ en | 67.53 | 71.67 | 59.99 |
| ar2→ en | 64.98 | 70.48 | 59.90 |

Table 3: Probing results of the [en] and English transferred models on the English-EWT test set using the structural probe model and on the WiC dev set.

- Semantic abstractions learned in the source language are transferred to English and enhance probing performance.
- Our cross-lingual method does not seem to boost the learning of syntactic information in the target language.

## Discussion

- We experimented only with the frequency-based initalisation of the embeddings, but future work will investigate other ways of mapping the source and target vocabulary spaces.
- We believe that our method can provide a smart initialization for training models in minority languages, where neither large amounts of data nor computational resources are available.

## Conclusion

In this paper, we show that using a pre-trained monoingual language model as initialization for pre-training in new language spaces is beneficial with regard to model performance on downstream tasks and linguistic knowledge in the target language.

## References

[1] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. "On the Cross-lingual Transferability of Monolingual Representations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

[2] Hyung Won Chung et al. "Rethinking Embedding Coupling in Pre-trained Language Models". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=xpFFI_NtgpW.

[3] Alexis Conneau et al. "Emerging Cross-lingual Structure in Pretrained Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

[4] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://www.aclweb.org/anthology/2020.acl-main.747.

[5] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

[6] John Hewitt and Christopher D. Manning. "A Structural Probe for Finding Syntax in Word Representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.

[7] K Karthikeyan et al. "Cross-Lingual Ability of Multilingual BERT: An Empirical Study". In: *International Conference on Learning Representations*. 2020.

[8] Mohammad Taher Pilehvar and Jose Camacho-Collados. "WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.

[9] Linting Xue et al. *mT5: A massively multilingual pre-trained text-to-text transformer*. arXiv:2010.11934. 2020. arXiv: 2010.11934 [cs.CL].

This work is supported by the Swedish innovation agency (Vinnova) under contract 2019-02996. We would like to thank Joakim Nivre for his useful feedback in this work and Fredrik Carlsson for the illustration.