Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 948–955
Marseille, 20-25 June 2022

Cross-lingual Transfer of Monolingual Models

Evangelia Gogoulou⋆, Ariel Ekgren◦, Tim Isbister◦, Magnus Sahlgren◦
⋆RISE Research Institutes of Sweden, ◦AI Sweden
evangelia.gogoulou@ri.se
{ariel.ekgren, tim.isbister, magnus.sahlgren}@ai.se

Abstract

Recent studies in cross-lingual learning using multilingual models have cast doubt on the previous hypothesis that shared vocabulary and joint pre-training are the keys to cross-lingual generalization. We introduce a method for transferring monolingual models to other languages through continuous pre-training and study the effects of such transfer from four different languages to English. Our experimental results on GLUE show that the transferred models outperform an English model trained from scratch, independently of the source language. After probing the model representations, we find that model knowledge from the source language enhances the learning of syntactic and semantic knowledge in English.

1. Introduction

Training a language model from scratch requires considerable resources, both with respect to data and computational resources. These requirements can be a limiting factor for many actors, and for smaller languages, it is not clear whether there even *exists* enough data to train a language model. Consequently, there is an increasing interest in using cross-lingual transfer to alleviate the requirements of training a language model from scratch. Most of the work in this direction (see the next section) concerns multilingual models.

In this paper, we want to explore the effects of cross-lingual transfer of a monolingual model into a target language space. More specifically, we want to investigate if it is possible to adapt existing monolingual models to the target language and study their downstream performance in comparison with a model trained from scratch in the target language. Additionally, we want to study the impact of language similarity between source and target language on fine-tuning performance after transfer. Based on recent work that shows that transformer-based language models encode universal properties (Lu et al., 2021), we hypothesize that model knowledge learned in the source language enhances the learning of the target language independently of language proximity.

Our contributions in this work are the following: **(i)** we introduce an adaptation method for cross-lingual transfer (Section 3.), **(ii)** we show that the models that have been transferred *to* English outperform an English model trained from scratch in the GLUE benchmark for *all* source languages studied here (Section 4.2.), **(iii)** through probing the model representations, we demonstrate that abstractions learned in the source language are transferred to English (Section 4.3.).

2. Related Work

The standard methodology of transferring knowledge across languages is by training either cross-lingual (Ruder et al., 2019) or multilingual models (Schwenk and Douze, 2017; Devlin et al., 2019; Conneau et al., 2020a; Xue et al., 2020; Chung et al., 2021). These latter models are trained on massively multilingual data using a shared vocabulary, which has proven to be successful for *zero-shot* cross-lingual transfer (Pires et al., 2019), where the multilingual model, in this case mBERT, is fine-tuned on a downstream task in the source language and evaluated on the same task in the target language.

Pires et al. (2019) hypothesize that zero-shot cross-lingual generalization is facilitated by using a shared vocabulary. Several recent studies contradict this assumption. Karthikeyan et al. (2020) show that in a joint-training setting, multilinguality can be achieved even if the two languages do not share any vocabulary. Conneau et al. (2020b) train jointly bilingual masked language models that share only the top two Transformer layers. A different perspective is provided by Artetxe et al. (2020), who disregard even the joint pre-training constraint and transfer monolingual BERT to a new language by learning only a new embedding matrix from scratch while freezing the rest of the model. Their results indicate that neither shared vocabulary nor joint pre-training are necessary for cross-lingual transfer in the zero-shot setting. This method has also been applied to GPT-2 (de Vries and Nissim, 2021).

The overarching conclusion mainly from Conneau et al. (2020b) and Artetxe et al. (2020) is that zero-shot cross-lingual transfer is facilitated by shared statistical properties between language spaces, rather than multilingual pre-training. We would like to explore if this holds in a different transfer scenario, where a model is transferred to a new language and fine-tuned on monolingual tasks in that language. Our hypothesis is that the statistics of language acquired by a model in the source language will transfer and boost monolingual task performance in the target language.

3. Method and Models

Due to the lack of standardized monolingual downstream tasks in non-English languages, we have chosen to transfer from other languages into English. However, our method is likely to be most useful in a low-resource

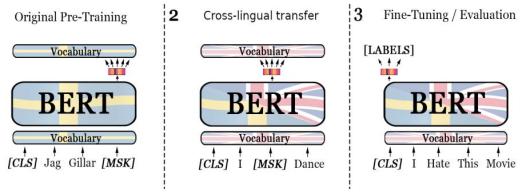

Figure 1: Our cross-lingual transfer method, applied to the transfer from Swedish to English: we continue pre-training Swedish BERT on our English Corpus, using the English vocabulary. Then, we fine-tune the transferred model on an English task.

scenario. The proposed method is illustrated in Figure 1. 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 and *continue* pre-training the model on English data. We denote a model transferred with our method from the source language *lang* to English by [*lang*→ en]. Given that the vocabulary tokens learned by a Wordpiece tokenizer are ordered by descending frequency, our method maps the vocabulary of the target language to the trained weights of the source embeddings with similar frequencies. In other words, the fifth most frequent token of the English vocabulary is initialised with the source embedding of the fifth most frequent token of the source vocabulary.

Our method is closely related to the MonoTrans method proposed by Artetxe et al. (2020). In both cases, cross-lingual transfer is primarily performed by adapting the embedding layer of a monolingual model. In (Artetxe et al., 2020), the embedding layer is learned from scratch in the target language, while the rest of the model parameters are frozen. In our case, we continue pre-training all model parameters in the target language, including the embedding layer. An additional difference is the method used for evaluating the model after transfer: we fine-tune the model on tasks in the target language, while Artetxe et al. (2020) perform zero-shot evaluation.

The main criteria for including a monolingual language model in our experiments is that its architecture and training procedure should follow BERT-base (Devlin et al., 2019) and its pre-training corpus should include Wikipedia. The complete list of monolingual BERT models employed is presented in Table 1, while the details of model selection can be found in Appendix section A. Note that we have selected languages with varying degrees of linguistic similarity to English, and also one language with two different models with different amounts of training data.

4. Experiments

4.1. Experimental Setup

Our pre-training corpus is English Wikipedia,[1] which amounts to 13G English text. We train a Wordpiece tokenizer on the downloaded English Wikipedia, similar to Devlin et al. (2019). The vocabulary size is fixed to 32K. For [*lang*] models with a larger vocabulary size, the vocabulary is resized to 32K by keeping only the first 32K vocabulary tokens.

Each [*lang*→ en] model is trained for one epoch on English Wikipedia using the masked language modeling training objective. The complete list of training hyperparameters is shared between all trained models and can be found in Appendix section B.

All models are fine-tuned on the tasks included in the GLUE benchmark (Wang et al., 2018), excluding WNLI given the known issues with the dataset construction.[2] The standard fine-tuning procedure is followed (Devlin et al., 2019) and the hyperparameters can be found in Appendix section B. We refrain from reporting the GLUE test results, since the comparison with models achieving SOTA performance in GLUE is not relevant for this work.

For each language pair (*lang*, en), the fine-tuning performance of model [*lang*→ en] is compared with the performance of a BERT-base model, namely [en], that is trained from scratch on our English corpus using the exact same setup with the transferred models. This baseline allows us to compare the effect of choosing different initializations, namely random or trained monolingual models, when training a model in a new language.

4.2. GLUE Fine-Tuning

The GLUE validation results of our models are reported in Table 2. The standard deviation of the metric score in each task is presented in Appendix section C. The results confirm that transferring a non-English model to English, namely [*lang*→ en], leads to significantly better

[1] Downloaded in November 2019, using this script: https://github.com/facebookresearch/XLM/blob/master/get-data-wiki.sh

[2] https://gluebenchmark.com/faq(Number 12)

Language	Model name	Alias	Vocab size	Data (GB)
English	BERT-base (ours)	en	32,000	13
Swedish	KB-BERT (Malmsten et al., 2020)	sv	50,325	18
Dutch	BERTje (de Vries et al., 2019)	nl	30,000	12
Finnish	FinBERT (Virtanen et al., 2019)	fi	50,105	≈ 48
Arabic	AraBERTv01 (Antoun et al., 2020)	ar1	64,000	23
	AraBERTv02 (Antoun et al., 2020)	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.

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: GLUE validation metric score for all models and tasks. Accuracy is the reported metric for all tasks, except from CoLA (Mathew correlation coefficient), QQP and MRPC (F1 score) and STS-B (Spearman correlation (x100)). The "AVG" column corresponds to the average score computed across all evaluated GLUE tasks. The bold font underlines the overall best performing model per source language, comparing to the [en] model.

fine-tuning performance than an English model trained from scratch, namely [en]. The best performing model, in terms of average score, is [sv→ en], outperforming [en] by 4.88 absolute difference. Additionally, we make the following observations:

- **All transferred models improve over the [en] model independently of source language:** Interestingly, the linguistic similarity between the source language and English does not significantly impact the effectiveness of our cross-lingual transfer method.
- **The data size of the pre-training corpus in the source language matters:** The English model transferred from [ar2], namely [ar2→ en], performs better than [ar1→ en], which originates from [ar1]. Given that [ar2] is trained on ≈ 3 times more data than [ar1], this possibly indicates the important role of pre-training data size in cross-lingual performance.

We also investigate the case where we apply English pre-training on the source monolingual models but *keep* the source vocabulary. The fine-tuning results on GLUE, presented in Table 3, demonstrate the importance of matching the tokenizer to the vocabulary of the target language. However, it is noteworthy that the models perform similarly (or even better) to the English model trained from scratch.

Lang	Vocab	AVG
en	en	72.97
sv→ en	sv	73.25
sv→ en	en	77.85
nl→ en	nl	72.99
nl→ en	en	76.13
fi→ en	fi	68.54
fi→ en	en	76.34
ar1→ en	ar1	73.99
ar1→ en	en	75.54
ar2→ en	ar2	74.27
ar2→ en	en	77.02

Table 3: Average GLUE validation score for all models, using the original or the English vocabulary. The bold font underlines the overall best performing model per source language.

4.3. English Linguistic Probing

We also study the linguistic effects of the proposed cross-lingual transfer method. This is done by evaluating the syntactic and semantic knowledge of the [en] compared to the [lang→ en] models through probing their representations.

More specifically, we evaluate the word representations yielded by [lang→ en] using the *structural probe* model, proposed by Hewitt and Manning (2019), which detects whether syntactic trees are encoded in a linear transformation of the model embedding space. In this way, we evaluate if the word embeddings of the transferred English models encode syntactic parsing information. Following Hewitt and Manning (2019), we define the structural probe model as a linear transformation that learns the tree distances between all pairs of words in training sentences from the English part of Universal Dependencies v2.7 (English-EWT).[3] The trained probing model is then evaluated on the English-EWT test set using the following evaluation metrics (Hewitt and Manning, 2019): Spearman correlation between predicted and true word pair distances (DSpr), averaged across the input sentences with length 5-50, and the percentage of undirected edges placed correctly in comparison with the gold parse tree, namely undirected unlabelled attachments score (UUAS).

For the semantic probing of the transferred models, we use the Words In Context task (WiC Pilehvar and Camacho-Collados (2019)). This is a binary classification task, where the model needs to determine if a given word is used with the same meaning or not in two different contexts. For this purpose, we train a linear classifier on top of each sentence representation. The details of our probing setup can be found in Appendix section D.

The probing results are presented in Table 4. The improvement of the [*lang*→ en] models over the [en]

[3] https://github.com/UniversalDependencies/UD_English-EWT

	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 4: Probing results of the [en] and English transferred models on the English-EWT test set using the structural probe model (Hewitt and Manning, 2019) and on the WiC (Pilehvar and Camacho-Collados, 2019) dev set.

model on the WiC test for all source languages demonstrates that semantic abstractions learned in the source language are transferred to English and enhance probing performance. Interestingly, the results of syntactic probing on the transferred models are in the best case similar to the probing results on the [en] model. It is worth observing that unlike the results on the GLUE benchmark, [ar2→ en] performs clearly worse than [ar1→ en] on syntactic probing. This indicates that larger pre-training data size hinders the learning of syntactic information in the target language through language pre-training. Overall, the results on English probing indicate that our cross-lingual transfer method boosts the learning of semantic information in the target language, but does not enhance the learning of syntactic information.

5. Discussion

The monolingual experimental setup employed here allows us to control for the source language and study the effect of this choice on cross-lingual performance. The presented results on GLUE as well as English probing tasks show that the linguistic similarity between source and target language is not important for cross-lingual transfer in our setup. This result contradicts previous work by Lauscher et al. (2020) that studies the effect of language similarity in zero-shot cross-lingual transfer and finds that multilingual language models have poor zero-shot performance in distant target languages.

Our setup differs significantly from *zero-shot* cross-lingual transfer, where the source model is transferred to the target language at test time, after being fine-tuned on a task in the source language. By contrast, we adapt trained monolingual models to the target language through additional pre-training. This is inspired by a domain adaptation approach proposed by Gururangan et al. (2020). They show that adapting the original language model to the target domain through extra pre-training improves model performance on the target task. Our results suggest that this approach is also beneficial for model transfer between two different language spaces. Future work will study the adaptation of multilingual models with our method and perform a comparison with zero-shot cross-lingual transfer.

As part of the proposed cross-lingual transfer method, each token embedding in the target embedding matrix is initialised with the trained source embedding at the same position in the source embedding matrix. The investigation of the effect of this frequency-based embedding initialisation scheme on model performance in the target language, in comparison with other types of initialisations such as random, is left for future work.

Due to the lack of established evaluation benchmarks in other languages, we perform cross-lingual transfer experiments only from other languages to English. However, 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. In this direction, de Vries et al. (2021) transfer monolingual BERT models to two Dutch dialects using zero-shot learning.

6. Conclusion

In this paper, we show that using a pre-trained 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. Our experimental results demonstrate that language similarity has no impact on cross-lingual performance, while larger pre-training data size appears to have a positive effect on monolingual task performance in the target language.

We hope that our work will inspire practitioners and researchers to initialize new monolingual models with parameters from existing models if possible, in order to reach competitive models in low resource settings and reach better performance on downstream tasks.

7. Acknowledgements

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 on this work and Fredrik Carlsson for the illustration of the method (Figure 1).

8. References

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Artetxe, M., Ruder, S., and Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 4623–4637, Online, July. Association for Computational Linguistics.

Chi, E. A., Hewitt, J., and Manning, C. D. (2020). Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online, July. Association for Computational Linguistics.

Chung, H. W., Fevry, T., Tsai, H., Johnson, M., and Ruder, S. (2021). Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020a). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., and Stoyanov, V. (2020b). Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online, July. Association for Computational Linguistics.

de Vries, W. and Nissim, M. (2021). As good as new. how to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online, August. Association for Computational Linguistics.

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., and Nissim, M. (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

de Vries, W., Bartelds, M., Nissim, M., and Wieling, M. (2021). Adapting monolingual models: Data can be scarce when language similarity is high. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4901–4907, Online, August. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Hewitt, J. and Manning, C. D. (2019). 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)*, pages 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Karthikeyan, K., Wang, Z., Mayhew, S., and Roth, D. (2020). Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, November. Association for Computational Linguistics.

Lu, K., Grover, A., Abbeel, P., and Mordatch, I. (2021). Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*.

Malmsten, M., Börjeson, L., and Haffenden, C. (2020). Playing with words at the national library of sweden–making a swedish bert. *arXiv preprint arXiv:2007.01658*.

Pilehvar, M. T. and Camacho-Collados, J. (2019). 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)*, pages 1267–1273, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Schwenk, H. and Douze, M. (2017). Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada, August. Association for Computational Linguistics.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP*

Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv:2010.11934.

Lang	CoLA (stdev)	MNLI (stdev) (m/mm)	MRPC (stdev)	QNLI (stdev)	QQP (stdev)	RTE (stdev)	SST-2 (stdev)	STS-B (stdev)
en	2.62	0.23/0.28	0.42	0.30	0.05	1.89	0.85	0.19
sv→ en	1.21	0.17/0.09	0.19	0.19	0.07	1.95	0.32	0.28
nl→ en	1.22	0.24/0.27	0.75	0.44	0.09	0.82	0.36	0.40
fi→ en	2.71	0.12/0.14	0.65	0.28	0.09	1.54	0.51	0.48
ar1→ en	2.63	0.16/0.18	0.71	0.54	0.10	2.87	0.68	0.15
ar2→ en	0.74	0.19/0.21	1.06	0.37	0.05	2.12	0.43	0.44

Table 5: Standard deviation in GLUE validation results. Each value corresponds to the square root of the sample variance from the mean, computed across 5 runs per combination of model and task.

A Selection of monolingual language models

All monolingual models used were downloaded from the HuggingFace model hub,[4] which is an open library. All selected models are cased, with the exception of AraBERTs. The mapping between alias name in the paper and model name in the Hugging Face model hub is presented in Table 6.

In Table 1, the pre-training data sizes of KB-BERT, BERTje and AraBERTs were taken directly from the corresponding papers (Malmsten et al., 2020; de Vries et al., 2019; Antoun et al., 2020). For BERT-base, our estimation of the pre-training data size is based on the total number of words (3.3B) in the pre-training corpus, provided by (Devlin et al., 2019). Our estimation for FinnBERT is based on the total number of characters (24B), stated in the original paper Virtanen et al. (2019).

Alias	Hugging Face name
BERT-base (original)	bert-base-cased
sv	KB/bert-base-swedish-cased
nl	GroNLP/bert-base-dutch-cased
fi	TurkuNLP/bert-base-finnish-cased-v1
ar1	aubmindlab/bert-base-arabertv01
ar2	aubmindlab/bert-base-arabertv02

Table 6: Mapping between model alias name in the paper and model name in the Hugging Face model hub.

B Training details

The implementation of English pre-training and GLUE fine-tuning is heavily based on the example scripts[5] provided by the HuggingFace `transformers` library. The hyperparameters used are presented in Table 7. The parameters which are not presented here have been set to their default value.[6] For training the English Wordpiece tokenizer, we used the example code which is part of the HuggingFace `tokenizers` library.[7] All models were trained on a single GPU machine (Nvidia Tesla v100 sxm2 32GB), with the average training time per model being 4 days.

Hyperparameter	Training value	Fine-tuning value
batch size	128	32
learning rate	5e-5	2e-5
maximum sequence length	128	128
train_epochs	1.0	3.0
optimizer	AdamW	AdamW

Table 7: Hyperparameters used for English pre-training (second column) and GLUE fine-tuning (third column).

C GLUE results

For each model and task, we perform 5 runs with varying random seeds. The standard deviation from the average performance across these 5 runs is presented in Table 5. Overall, we observe that the standard deviation is notably higher in CoLA and RTE comparing to the rest of the GLUE tasks.

D English linguistic probing experiments

D1. Syntactic probing

Our implementation of the structural probing method (Hewitt and Manning, 2019) is heavily based on the coding repositories[8] of (Hewitt and Manning, 2019; Chi et al., 2020).

D2. Semantic probing

The `jiant` library[9] was used for training and evaluation of our models on the WiC task (Pilehvar and Camacho-Collados, 2019). The reported accuracy corresponds to the average accuracy across 5 runs, each one with a different random seed. The standard deviation across runs is presented in Table D2..

[4] https://huggingface.co/

[5] https://github.com/huggingface/transformers/blob/master/examples/pytorch/language-modeling/run_mlm.py, https://github.com/huggingface/transformers/blob/master/examples/pytorch/text-classification/run_glue.py

[6] https://huggingface.co/transformers/_modules/transformers/training_args.html

[7] https://huggingface.co/docs/tokenizers/python/latest/pipeline.html#all-together-a-bert-tokenizer-from-scratch

[8] https://github.com/ethanachi/multilingual-probing-visualization https://github.com/john-hewitt/structural-probes/

[9] https://github.com/nyu-mll/jiant

Lang	WiC (stdev)
en	1.19
sv→ en	1.27
nl→ en	0.76
fi→ en	1.76
ar1→ en	1.10
ar2→ en	1.22

Table 8: Standard deviation (stdev) of the average accuracy on WiC validation set over 5 runs with different random seeds. Standard deviation is computed as the square root of the sample variance from the mean accuracy.