Evaluating Transferability of BERT Models on Uralic Languages

Judit Ács
SZTAKI
Institute for Computer Science and Control
judit@sch.bme.hu

Dániel Lévai
Department of Digital Humanities
Eötvös Loránd University
levai753@gmail.com

András Kornai
SZTAKI
Institute for Computer Science and Control
kornai@sztaki.hu

Abstract

Transformer-based language models such as BERT have outperformed previous models on a large number of English benchmarks, but their evaluation is often limited to English or a small number of well-resourced languages. In this work, we evaluate monolingual, multilingual, and randomly initialized language models from the BERT family on a variety of Uralic languages including Estonian, Finnish, Hungarian, Erzya, Moksha, Karelian, Livvi, Komi Permyak, Komi Zyrian, Northern Sámi, and Skolt Sámi. When monolingual models are available (currently only et, fi, hu), these perform better on their native language, but in general they transfer worse than multilingual models or models of genetically unrelated languages that share the same character set. Remarkably, straightforward transfer of high-resource models, even without special efforts toward hyperparameter optimization, yields what appear to be state of the art POS and NER tools for the minority Uralic languages where there is sufficient data for finetuning.

A BERT- és más Transformer-alapú nyelvmodellek számos angol tesztadaton jobban teljesítenek, mint a korábbi modellek, azonban ezek a tesztadatok az angolra és néhány hasonlóan sok erőforrással rendelkező nyelvre korlátozódnak. Ebben a cikkben egynyelvű, soknyelvű és random súlyokkal inicializált BERT modelleket értékelünk ki a következő uráli nyelvekre: észt, finn, magyar, erza, moksa, karjalai, livvi-karjalai, komi-permják, komi-zürjén, északi számi és kolta számi. Az egynyelvű modellek – jelenleg csak észt, finn és magyar érhető el – ugyan jobban teljesítenek az adott nyelvre, általában rosszabbul transzferálhatóak, mint a soknyelvű modellek vagy a nem rokon, de azonos írást használó egynyelvű modellek. Érdekes módon a sok erőforráson tanult modellek még hiperparaméter optimalizálás nélkül is könnyen transzferálhatók és finomhangolásra alkalmas tanítóadattal csúcsminőségű POS és NER taggerek hozhatóak létre a kisebbségi uráli nyelvekre.

1 Introduction

Contextualized language models such as BERT (Devlin et al., 2019) drastically improved the state of the art for a multitude of natural language processing applications. Devlin et al. (2019) originally released 4 English and 2 multilingual pretrained versions of BERT (mBERT for short) that support over 100 languages including three Uralic languages: Estonian [et], Finnish [fi], and Hungarian [hu]. BERT was quickly followed by other large pretrained Transformer (Vaswani et al., 2017) based models such as RoBERTa (Liu et al., 2019) and multilingual models such as XLM-RoBERTa (Conneau et al., 2019). Huggingface released the Transformers library (Wolf et al., 2020), a PyTorch implementation of Transformer-based language models along with a repository for pretrained models from community contribution[1]. This list now contains over 1000 entries, many of which are domain-specific or monolingual models.

Despite the wealth of multilingual and monolingual models, most evaluation methods are limited to English, especially for the early models. Devlin et al. (2019) showed that the original mBERT outperformed existing models on the XNLI dataset (Conneau et al., 2018), a translation

---

[1] https://huggingface.co/models

of the MultiNLI (Williams et al., 2018) to 15 languages. mBERT was further evaluated by Wu and Dredze (2019) for 5 tasks in 39 languages, which they later expanded to over 50 languages for part-of-speech (POS) tagging, named entity recognition (NER) and dependency parsing (Wu and Dredze, 2020). mBERT has been applied to a variety of multilingual tasks such as dependency (Kondratyuk and Straka, 2019) and constituency parsing (Kitaev et al., 2019). The surprisingly effective multilinguality of mBERT was further explored by Dufter and Schütze (2020).

Uralic languages have received relatively moderate interest from the language modeling community. Aside from the three national languages, no other Uralic language is supported by any of the multilingual models, nor does any have a monolingual model. There are no Uralic languages among the 15 languages of XNLI. Wu and Dredze (2020) do explore all 100 languages that mBERT supports but do not go into monolingual details. Alnajjar (2021) transfer existing BERT models to minority Uralic languages, the only work that focuses solely on Uralic languages.

In this paper we evaluate multilingual and monolingual models on Uralic languages. We consider three evaluation tasks: morphological probing, POS tagging and NER. We also use the models in a crosslingual setting, in other words, we test how monolingual models perform on related languages. We show that

- these language models are very good at all three tasks when finetuned on a small amount of task specific data,

- for morphological tasks, when native BERT models are available (et, fi, hu), these outperform the others on their native language, though the advantage over XLM-RoBERTa is not statistically significant,

- for POS and NER, the use of native models from related, even closely related languages, rarely brings improvement over the multilingual models or even English models,

- as long as the alphabet that the language uses is covered in the vocabulary of the model, we can transfer mBERT (or RuBERT) to the NER and POS tasks with surprisingly little finetuning data.

2 Approach

We evaluate the models through three tasks: morphological probing, POS tagging and NER. Uralic languages have rich inflectional morphology and largely free word order. Morphology plays a key role in parsing sentences. Morphological probing tries to recover morphological tags from the sentence representation from these models.

For assessing the sentence level behavior of the models we chose two token-level sentence tagging tasks, POS and NER. Part of speech tagging is a common subtask of downstream NLP applications such as dependency parsing. Named entity recognition is indispensable for various high level semantic applications such as building knowledge graphs. Our model architecture is identical for POS and NER.

2.1 Morphological probing

Probing is a popular evaluation method for black box models. Our approach is illustrated in Figure 1. The input of a probing classifier is a sentence and a target position (a token in the sentence). We feed the sentence to the contextualized model and extract the representation corresponding to the target token. Early experiments showed that lower layers retain more morphological information than higher layers so instead of using the top layer, we take the weighted average of all Transformer layers and the embedding layer. The layer weights are learned along with the other parameters of the neural network. We train a small classifier on top of this representation that predicts a morphological tag. We expose the classifier to a limited amount of training data (2000 training and 200 validation instances). If the classifier performs well on unseen data, we conclude that the representation includes the relevant morphological information.

We generate the probing data for Estonian and Finnish from the Universal Dependencies (UD) Treebanks (Nivre et al., 2020; Haverinen et al., 2014; Pyysalo et al., 2015; Vincze et al., 2010) and from the automatically tagged Webcorpus 2.0 for Hungarian since the Hungarian UD is very small. Unfortunately we could not extend the list of languages to other Uralic languages because their treebanks are too small to sample enough data.

The sampling method is constrained so that the target words have no overlap between train, validation and test, and we limit class imbalance to 3-to-1 which resulted in filtering some rare values. We
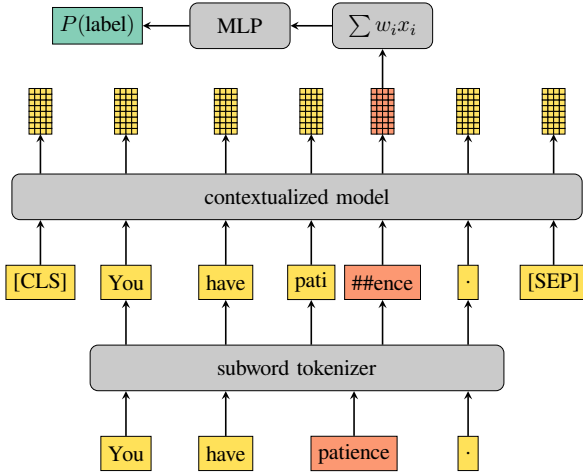
Figure 1: Probing architecture. Input is tokenized into subwords and a weighted average of the mBERT layers taken on the last subword of the target word is used for classification by an MLP. Only the MLP parameters and the layer weights w_i are trained.

were able to generate enough probing data for 11 Estonian, 16 Finnish and 11 Hungarian tasks, see Table 4 for the full list of these.

2.2 Sequence tagging tasks

Our setup for the two sequence tagging tasks is similar to that of the morphological probes except we train a shared classifier on top of all token representations. We use the vector corresponding to the first subword in both tasks. Although this may be suboptimal in morphology, Ács et al. (2021) showed that the difference is smaller for POS and NER. We also finetune the models which seems to close the gap between first and last subword pooling for morphology, see 4.1. For sequence tagging tasks, unlike for morphology, we found that the weighted average of all layers is suboptimal compared to simply using the top layer, so the experiments presented here all use the top layer.

We sample 2000 train, 200 validation and 200 test sentences as POS training data from the largest UD treebank in Estonian and Finnish, and from Webcorpus 2.0 for Hungarian. Aside from these three, Erzya [myv]; Moksha [mdf]; Karelian [krl]; Livvi [olo]; Komi Permyak [koi]; Komi Zyrian [kpv]; Northern Sámi [sme]; and Skolt Sámi [sms] have UD treebanks (Rueter and Tyers, 2018; Rueter, 2018; Pirinen, 2019; Rueter, 2014; Rueter et al., 2020; Partanen et al., 2018; Sheyanova and Tyers, 2017), but these are considerably smaller in size.

Language	Code	Morph	POS	NER
Hungarian	[hu]	26k	2000	2000
Finnish	[fi]	38k	2000	2000
Estonian	[et]	26k	2000	2000
Erzya	[myv]	0	1680	1800
Moksha	[mdf]	0	164	400
Karelian	[krl]	0	224	0
Livvi	[olo]	0	122	0
Komi Permyak	[koi]	0	78	2000
Komi Zyrian	[kpv]	0	562	1700
Northern Sámi	[sme]	0	2000	1200
Skolt Sámi	[sms]	0	101	0

Table 1: Size of training data for each language.

Although none of these languages are officially supported by any of the language models we evaluate, we train crosslingual models and find that the models have remarkable crosslingual capabilities.

Our NER data is sampled from WikiAnn (Pan et al., 2017). WikiAnn has data in Erzya, Estonian, Finnish, Hungarian, Komi Permyak, Komi Zyrian, Moksha, and Northern Sámi.[2] Similarly to the POS training data, we sample 2000 training, 200 validation and 200 test sentences when available, see Table 1 for actual training set sizes.

2.3 Training details

We train all classifiers with identical hyperparameters. The classifiers have one hidden layer with 50 neurons and ReLU activation. The input and the output dimensions are determined by the choice of language model and the number of target labels. The classifiers have 40 to 60k trainable parameters which are randomly initialized and updated using the backpropagation algorithm. We run experiments both with and without finetuning the language models. Finetuning involves updating both the language model (all 110M parameters) and the classification layer (end-to-end training).

All models are trained using the AdamW optimizer (Loshchilov and Hutter, 2019) with $lr = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999$. We use 0.2 dropout for regularization and early stopping based on the development set. We set the batch size to 128 when not finetuning the models, and we use batch size 8, 12 or 20 when we finetune them.

The evaluated models, all from the

---

[2]WikiAnn also has Udmurt data, but the transcription is problematic: Latin and Cyrillic are used inconsistently, Wikipedia Markup is parsed incorrectly etc.

BERT/RoBERTa family, differ only in the choice of training data and the training objective. They all have 12 Transformer layers, with 12 heads, and 768 hidden dimensions, for a total of 110M parameters.

3 The models evaluated

Our goal is twofold: we want to assess monolingual models against multilingual models, and we want to evaluate the models on 'unsupported' languages, both typologically related and unrelated.

We pick two multilingual models, mBERT and XLM-RoBERTa. Our choices for monolingual models are EstBERT for Estonian, FinBERT for Finnish and HuBERT for Hungarian (See Table 2). As a control, we also test the English BERT as a general test for cross-language transfer. Since many Uralic speaking communities are in Russia and the languages are heavily influenced by Russian, we test RuBERT on these languages. Finally, we also test a randomly initialized mBERT. We do this because the capacity of the BERT-base models is so large that they may memorize the probing data alone. Many models have cased and uncased version, the latter often removing diacritics along with lowercasing. Since diacritics play an important role in many Uralic languages, we only use the cased models. We return to this issue in 3.1.

The models along with their string identifier are summarized in Table 2.

3.1 Subword tokenization

Subword tokenization is a key component in achieving good performance on morphologically rich languages. There are two different tokenization methods used in the models we compare: XLM-RoBERTa uses the SentencePiece algorithm (Kudo and Richardson, 2018), the other models use the WordPiece algorithm (Schuster and Naka-jima, 2012). The two types of tokenizers are algorithmically very similar, the differences between them are mainly dependent on the vocabulary size per language. The multilingual models consist of about 100 languages, and the vocabularies per language apper sublinearly proportional to the amount of training data available per language: in case of mBERT, 77% of the word pieces are pure ascii (Ács, 2019).

The native models, trained on monolingual data, have longer and more meaningful subwords (see the bolded entries in Table 3). This greatly facilitates the sharing of train data, a matter of great importance for Uralic languages where there is little text available to begin with.

Both BERT- and RoBERTa-based models first tokenize along whitespaces, but the handling of missing characters differs significantly. In BERT-based models, if there is a character missing from the tokenizer's vocabulary, the model discards the whole segment between whitespaces, labeling it [UNK]. In cross-lingual cases many words are lost since monolingual models tend to lack the extra characters of a different language. In contrast, XLM-RoBERTa deletes the unknown characters, but the string that remains between whitespaces is segmented, so the loss of information is not as severe.

Table 3 summarizes different measures in language-model pairs. As a general observation, Latin script models (FinBERT, HuBERT, Est-BERT) are unusable on Cyrillic text, as seen e.g. on Erzya, where Latin script models produce [UNK] token for 97.5% of the word types. This is also seen for Northern Sámi and Hungarian, which have many non-ascii characters (á, é, í, ó, ö, ő, ú, ü, ű for Hungarian, č, đ, ŋ, š, ŧ, ž for Northern Sámi) see the Hungarian-EstBert/FinBERT pairs and the Northern Sámi-FinBERT/HuBERT pairs.

The mean subword length generally lies between 3.0 and 3.5 for most pairs - naturally, the corresponding language-model pairs have much higher mean subword length, 5.0 to even 5.9. This range is true not only for Latin script languages, but for Cyrillic script languages as well, as indicated by Erzya, which has a mean subword length of 3.1 to 3.4 on the multilingual models and on RuBERT.

Fertility (Ács, 2019) is defined as the average number of BERT word pieces found in a single real word type. EstBERT on Estonian and FinBERT on Finnish have very similar fertility values (2.1 and 1.9), but HuBERT on Hungarian has much higher fertility. This is mainly caused by the different vocabulary sizes - the Finnic models have 50000 subwords in their vocabulary, HuBERT only contains 32000 subwords. The rest of the fertility values are mostly over 3. In extreme cases, a word is segmented into letters, which is the case for EngBERT on Erzya, but the non-Hungarian models on Hungarian also produce very high fertility values.

Model	Identifier	Language(s)	Reference
mBERT	bert-base-multilingual-cased	100+ inc. et, fi, hu	Devlin et al. (2019)
XLM-RoBERTa	xlm-roberta-base	100 inc. et, fi, hu	Liu et al. (2019)
EstBERT	tartuNLP/EstBERT	Estonian	Tanvir et al. (2021)
FinBERT	TurkuNLP/bert-base-finnish-cased-v1	Finnish	Virtanen et al. (2019)
HuBERT	SZTAKI-HLT/hubert-base-cc	Hungarian	Nemeskey (2020)
EngBERT	bert-base-cased	English	Devlin et al. (2019)
RuBERT	DeepPavlov/rubert-base-cased	Russian	Kuratov and Arkhipov (2019)
rand-mBERT	mBERT with random weights	any	described in Section 3

Table 2: List of models we evaluate.

	mBERT	RoBERTa	EstBERT	FinBERT	HuBERT	RuBERT	EngBERT
Vocab. size	120k	250k	50k	50k	32k	120k	29k
Missing [et] (%)	.0	.0	**.2**	.0	.5	.1	.2
Missing [fi] (%)	.0	.0	.0	**.0**	.4	.0	.0
Missing [hu] (%)	.1	.0	21.5	48.3	**.1**	2.7	.2
Missing [sme] (%)	.2	.0	15.0	47.4	5.1	4.8	.2
Missing [myv] (%)	.0	.0	97.5	97.5	97.5	.0	.0
Subword length [et]	3.7±1.4	4.2±1.7	**5.8**±2.6	3.7±1.4	3.1±1.2	3.1±1.2	3.5±1.4
Subword length [fi]	3.8±1.4	4.5±1.9	3.8±1.4	**5.9**±2.5	3.1±1.1	3.1±1.1	3.4±1.4
Subword length [hu]	3.5±1.5	4.2±2.0	3.3±1.2	3.1±1.1	**5.0**±2.4	3.0±1.1	3.3±1.4
Subword length [sme]	3.2±1.0	3.4±1.1	3.2±1.1	3.2±1.1	3.1±1.2	2.9±1.0	3.0±1.0
Subword length [myv]	3.1±1.2	3.2±1.0	1.0±0.0	1.0±0.0	1.0±0.0	3.4±1.2	1.1±0.4
Character length [et]	9.2	9.2	**9.2**	9.2	9.2	9.2	9.2
Character length [fi]	9.3	9.3	9.3	**9.3**	9.3	9.3	9.3
Character length [hu]	9.8	9.8	9.6	8.8	**9.8**	9.8	9.9
Character length [sme]	8.5	8.5	8.3	7.6	8.5	8.4	8.5
Character length [myv]	7.3	7.3	1.8	1.8	1.7	7.3	7.3
Fertility [et]	3.4	2.8	**2.1**	3.6	4.4	4.3	4.3
Fertility [fi]	3.3	2.7	3.5	**1.9**	4.6	4.4	4.5
Fertility [hu]	4.0	3.2	5.2	4.5	**2.8**	5.4	5.6
Fertility [sme]	3.7	3.6	4.1	3.3	4.5	4.6	4.7
Fertility [myv]	3.6	3.3	1.1	1.1	1.1	3.0	7.2

Table 3: Major characteristics of cross-language tokenization. Boldface font marks the corresponding language-model pairs.
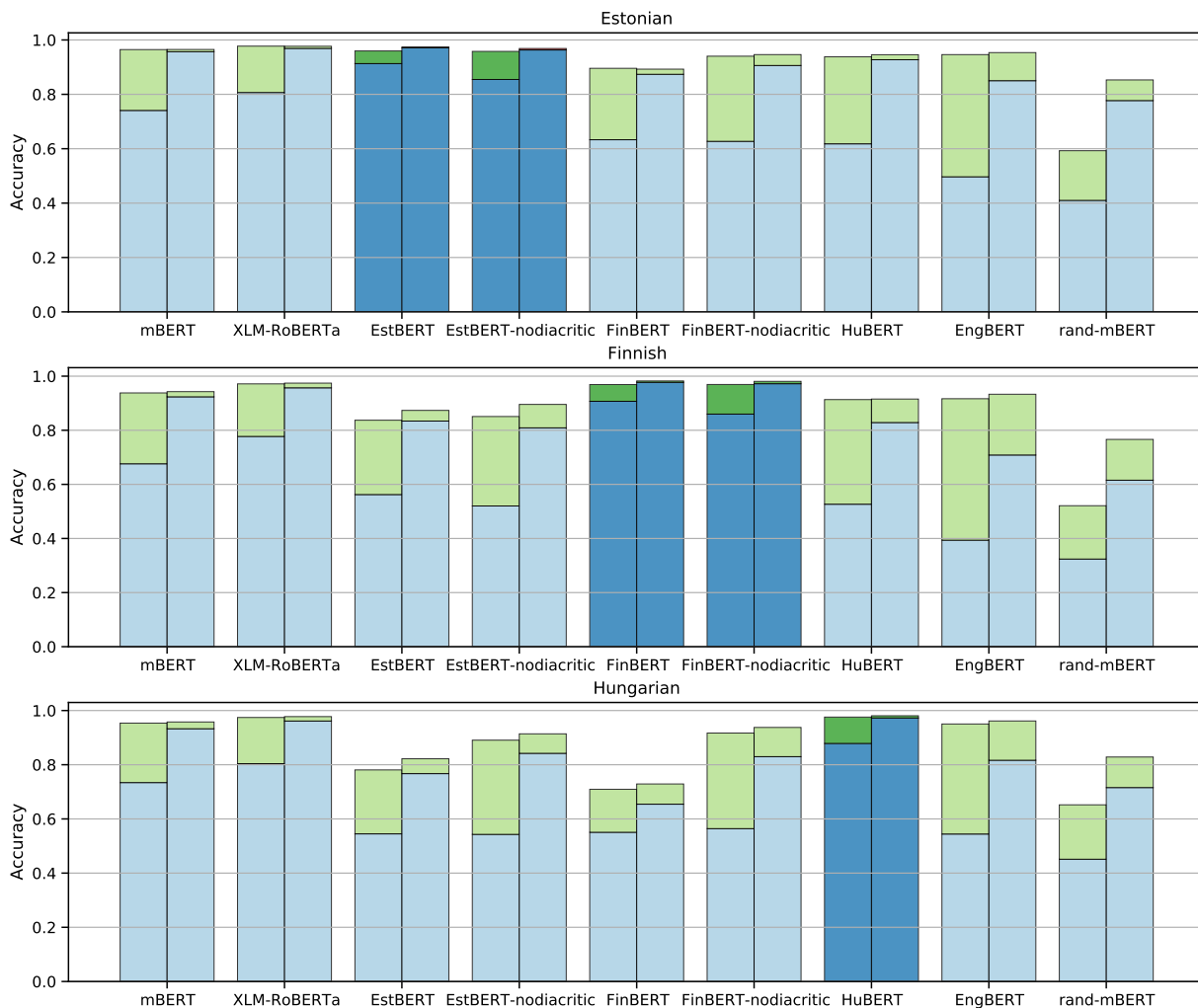
Figure 2: Mean accuracy of morphological tasks by language. The bars are grouped in two, the left one is the result of probing the first subword, the right one is the results of probing the last subword. Blue bars are without finetuning, green bars are with finetuning. Monolingual models are highlighted.

4 Results

4.1 Morphology

Morphological tasks are generally easy for most models and we see reasonable accuracy from crosslingual models as illustrated by Figure 2. Mean accuracies, especially after finetuning, are generally above 90%, except, unsurprisingly, for the randomly initialized models.

Subword choice We first start by examining the choice of subword on morphological tasks. We try probing the first and the last subword and we find that there is a substantial gap in favor of the last subword. This is unsurprising considering that Uralic languages are mainly suffixing. This gap on average shrinks from 0.21 to 0.032 when we finetune the models on the probing data (Figure 2 shows this gap in green). Without finetuning there is only one

task, ⟨Hungarian, Degree, ADJ⟩, where probing the first subword is better than probing the last one for some models. This is explained by the fact that the superlative in Hungarian is formed from the comparative by a prefix.

Monolingual models are only slightly better than the two multilingual models, XLM-RoBERTa in particular. We run paired t-tests on the accuracy of each model pair over the 11 (et, hu) or 16 (fi) morphological tasks in a particular language and find that the difference between the monolingual model and XLM-RoBERTa is never significant, and for Estonian, neither is the difference between Est-BERT and mBERT.

Cross-lingual transfer works only if we finetune the models. Interestingly, language relatedness does not seem to play a role here. FinBERT transfers

worse to Estonian than HuBERT, and EstBERT transfers worse to Finnish than HuBERT. Interestingly, EngBERT transfers better to all three models than the other native BERTs, and for Finnish and Hungarian it is actually on par with mBERT.

Diacritics As seen from the first panel of Table 3, EstBERT and FinBERT replace words with unknown characters with [UNK] to such an extent that a large proportion of types end up being filtered. We try to mitigate this issue by preemptively removing all diacritics from the input. It appears that this has little effect on the original language, but cross-lingual transfer is improved for Finnish. In the sequence tagging tasks that we now turn to, we remove the diacritics when we evaluate EstBERT or FinBERT in a cross-lingual setting.
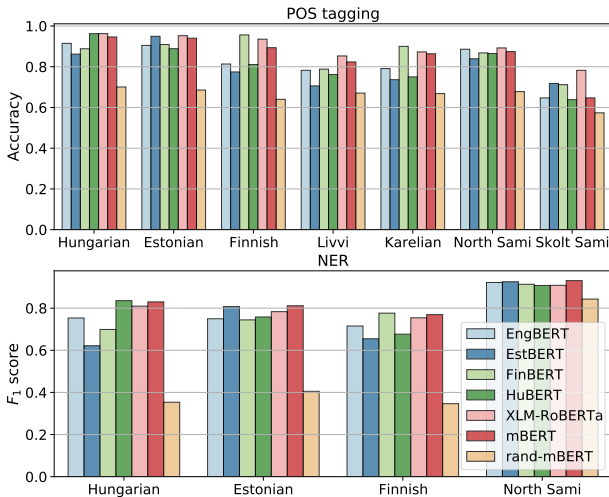
4.2 POS and NER

Figure 3: POS and NER results on languages that use the Latin alphabet.

We extend our studies to all Uralic languages with any training data (see Table 1) and we limit the discussion to finetuned models since cross-lingual transfer does not work without finetuning. We split the languages into two groups, Latin and Cyrillic, and we only test models with explicit support for the script that the language uses. Multilingual models support both scripts. Figures 3 and 4 show the results by language.

National languages We generally find the best performance in the three languages with native support: Estonian, Finnish and Hungarian. Monolingual models perform the best in their respective language but the two multilingual models are also very capable.
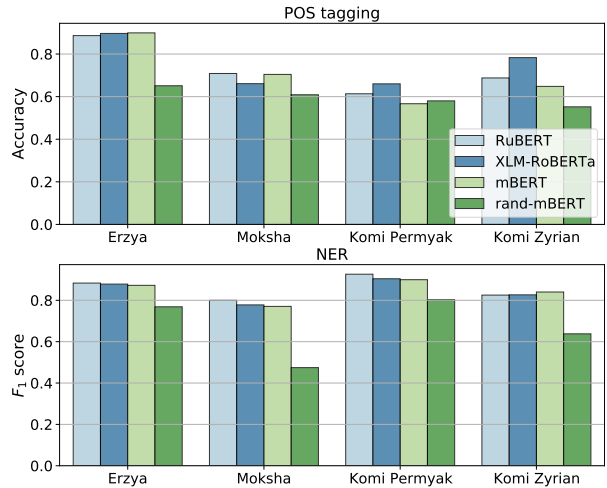

Figure 4: POS and NER results on languages that use the Cyrillic alphabet.

Cross-lingual transfer does not seem to benefit from language relatedness, EngBERT transfers just as well as other monolingual models. Even extremely close relatives such as Livvi and Finnish do not transfer better than XLM-RoBERTa to Livvi. On the other hand, FinBERT is the best for Karelian POS, another close relative of Finnish. The writing system and shared vocabulary also seem to play an important role, as seen from RuBERT's usefulness on unrelated but Cyrillic-using Uralic languages, see Figure 4.

XLM-RoBERTa is generally a strong model for cross-lingual transfer for all Uralic languages. We suspect that this is due to its large subword vocabulary, which may provide a better generalization basis for capturing the orthographic cues that are often highly indicative in agglutinative languages.

North Sámi Both POS and NER in North Sámi are relatively easy as long as the orthographic cues can be captured (i.e. the Latin script is supported). rand-mBERT is suprisingly successful at NER in North Sámi, suggesting that orthograpic cues (rand-mBERT uses mBERT's tokenizer) are highly predictive of named entities in North Sámi.

5 Conclusion

Altogether we find that it is possible, and relatively easy, to transfer models to new languages with finetuning on very limited training data, though extremely limited data still hinders progress: compare Erzya (1680 train sentences) to Moksha (164 train sentences) on Fig. 4.

Morph tag	POS	Estonian	Finnish	Hungarian
Case	adj	8 classes	11 classes	
Case	noun	15 classes	12 classes	18 classes
Case	propn		8 classes	
Case	verb		12 classes	
Degree	adj		Cmp, Pos, Sup	Cmp, Pos, Sup
Derivation	adj		Inen, Lainen, Llinen, Ton	
Derivation	noun		Ja, Lainen, Minen, U, Vs	
InfForm	verb		1, 2, 3	
Mood	verb		Cnd, Imp, Ind	Cnd, Imp, Ind, Pot
Number psor	noun			Sing, Plur
Number	a/n/v	Sing, Plur	Sing, Plur	Sing, Plur
PartForm	verb		Pres, Past, Agt	
Person psor	noun			1, 2, 3
Person	verb	1, 2, 3		1, 2, 3
Tense	adj	Pres, Past		
Tense	verb	Pres, Past	Pres, Past	Pres, Past
VerbForm	verb	Conv, Fin, Inf, Part, Sup	Inf, Fin, Part	Inf, Fin
Voice	adj	Act, Pass		
Voice	verb	Act, Pass	Act, Pass	

Table 4: List of morphological probing tasks.

EngBERT and RuBERT, which we introduced as a control for language transfer among genetically unrelated languages, transfer quite well: in particular the Latin-script EngBERT transfers better to Hungarian than FinBERT or EstBERT.

We note that we did not perform monolingual hyperparameter search or any preprocessing, and there is probably room for improvement for each of these languages. The biggest immediate gains are expected from extending the UD and WikiAnn datasets, and from careful handling of low-level characterset and subword tokenization issues. There are many Uralic languages that still lack basic resources, in particular the entire Samoyedic branch, Mari, and Ob-Ugric languages, are currently out of scope. Another avenue of research could be to work towards a stronger mBERT interlingua, or perhaps one for each script family, as the charset issues are clearly relevant.

Our data, code and the full result tables will be available along with the final submission.

References

Judit Ács, Ákos Kádár, and Andras Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Khalid Alnajjar. 2021. When word embeddings become endangered. *Multilingual Facilitation*, page 275–288.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk,

and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531. Open access.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Dávid Márk Nemeskey. 2020. *Natural Language Processing Methods for Language Modeling*. Ph.D. thesis, Eötvös Loránd University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132, Brussels, Belgium. Association for Computational Linguistics.

Tommi A Pirinen. 2019. Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 132–136.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of NoDaLiDa 2015*, pages 163–172. NEALT.

Jack Rueter. 2014. The Livonian-Estonian-Latvian Dictionary as a threshold to the era of language technological applications. *Journal of Estonian and Finno-Ugric Linguistics*, 5(1):251–259. ESUKA – JEFUL 2013, 5–1: 253–261 Volume: Proceeding volume:.

Jack Rueter. 2018. Erme ud moksha.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2020. On the questions in developing computational infrastructure for Komi-permyak. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 15–25, Wien, Austria. Association for Computational Linguistics.

Jack Rueter and Francis Tyers. 2018. Towards an open-source universal-dependency treebank for erzya. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 106–118.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Mariya Sheyanova and Francis M. Tyers. 2017. Annotation schemes in north sámi dependency parsing. In *Proceedings of the 3rd International Workshop for Computational Linguistics of Uralic Languages*, pages 66–75.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. Estbert: A pretrained language-specific bert for estonian.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, and S. Pyysalo. 2019. Multilingual is not enough: BERT for Finnish.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Judit Ács. 2019. Exploring bert's vocabulary. http://juditacs.github.io/2019/02/19/bert-tokenization-stats.html. Accessed: 2021-05-14.