WikiBERT Models: Deep Transfer Learning for Many Languages

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, Filip Ginter

TurkuNLP group,

Department of Computing, Faculty of Technology University of Turku, Finland first.last@utu.fi

Abstract

Deep neural language models such as BERT have enabled substantial advances in natural language processing. However, due to the effort and computational cost involved in their pre-training, such models are typically introduced only for highresource languages. In this paper, we introduce a simple, fully automated pipeline for creating language-specific BERT models from Wikipedia data and introduce 42 new monolingual models, most for languages up to now lacking such resources. We show that the newly introduced Wiki-BERT models outperform multilingual BERT (mBERT) in cloze tests for nearly all languages, and that parsing using Wiki-BERT models outperforms mBERT on average, with substantially improved performance for some languages, but decreases for others. All of the resources introduced in this work are available under open licenses from https://github.com/ turkunlp/wikibert.

1 Introduction

Transfer learning using language models pretrained on large unannotated corpora has allowed for substantial recent advances at a broad range of natural language processing (NLP) tasks. By contrast to earlier distributional semantics approaches such as random indexing (Kanerva et al., 2000) and context-independent neural approaches such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), models such as ULMFiT (Howard and Ruder, 2018), ELMo (Peters et al., 2018), GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) create contextualized representations of meaning, capable of providing both contextualized word embeddings as well as embeddings for text segments longer than words. Recent pre-trained neural language models have been rapidly advancing the state of the art in a range of natural language understanding and NLP tasks (Wang et al., 2018, 2019; Straková et al., 2019; Kondratyuk and Straka, 2019).

The transformer architecture (Vaswani et al., 2017) and the BERT language model of Devlin et al. (2019) have been particularly influential, with transformer-based models in general and BERT in particular fuelling a broad range of advances and serving as the basis of many recent studies of neural language models (e.g. Lan et al., 2019; Liu et al., 2019; Sanh et al., 2019). As is the case for most studies on new deep neural language models, the original study introducing BERT addressed only English. The authors later released a Chinese model as well as a multilingual model, mBERT, trained on text from 104 languages, but opted not to introduce models specifically targeting other languages. While mBERT is a powerful multilingual model with remarkable cross-lingual capabilities (Pires et al., 2019), it remains a compromise in that the 104 languages share the model capacity dedicated to one language in monolingual models, and it consequently suffers from degradation of performance in language-specific tasks (Conneau et al., 2020).

Here, we take steps towards closing various parts of the gap between languages with dedicated deep neural models, ones that share capacity with others in a massively multilingual model, and ones that lack any representation at all. We introduce a fully automated pipeline for creating languagespecific BERT models from Wikipedia data and apply this pipeline to create 42 new such models.

2 Related work

Considerable recent effort by various groups has focused on introducing dedicated BERT models covering single languages or a small number of (often closely related) languages. Dedicated monolingual models include e.g. BERTje¹ (de Vries et al., 2019) for Dutch, CamemBERT² (Martin et al., 2020) for French, FinBERT³ (Virtanen et al., 2019) for Finnish, RuBERT⁴ (Kuratov and Arkhipov, 2019) for Russian, and Romanian BERT (Dumitrescu et al., 2020); more focused multilingual models include e.g. the bilingual Finnish-English model of Chang et al. (2020) and the trilingual Finnish-Estonian-English and Croatian-Slovenian-English models of Ulčar and Robnik-Šikonja (2020).

Many of these studies have demonstrated the newly introduced models to allow for substantial improvements over mBERT in various languagespecific downstream task evaluations, thus supporting the continued value of creating monolingual and focused multilingual models. However, these efforts still cover only a fairly limited number of languages, and do not offer a straightforward way to substantially extend that coverage. The studies further differ considerably in aspects such as data collection, text cleaning and preprocessing, pre-training parameter setting and other details of the pre-training process, making it difficult to meaningfully compare the models to address questions such as which languages benefit most from mono/multilingual pre-training? We are not aware of previous efforts to automate the creation of large numbers of monolingual deep neural language models from comparable, publicly available sources nor efforts to create broadcoverage collections of such models.

In a line of study in some senses orthogonal to our work, a number of massively multilingual models improving on mBERT in terms of model architecture, training dataset, objectives, and process or other aspects have been introduced (e.g. Conneau et al., 2020; Xue et al., 2020). While it is certainly an interesting question to ask what the tradeoffs between monolingual and massively multilingual pre-training are for models other than BERT, it is not feasible for us to replicate the training processes for other models, and we have here chosen to focus on BERT-based models and Wikipedia due to their prominence and status as benchmarks.

```
deeppavlov/
```

3 Data

We next introduce the two primary datasets used in this study: Wikipedia, used as the source of unannotated texts for model pre-training, and Universal Dependencies annotated corpora, used to train preprocessing methods as well as in model evaluation.

3.1 Wikipedia

Wikipedia is a collaboratively created online encyclopedia that is available in a large number of languages under open data licenses. The English Wikipedia was the main source of text for pretraining the original English BERT models, accounting for three-fourths of its pre-training data.⁵ The mBERT models were likewise trained exclusively on Wikipedia data. In this work, we chose to focus on the Wikipedias in various languages as the only source of pre-training data, thus assuring that our approach can be directly applied to a broad selection of languages and providing direct comparability with existing models, in particular mBERT.

As of this writing, the List of Wikipedias⁶ identifies Wikipedias in 309 languages. Their sizes vary widely: while the largest of the set, the English Wikipedia, contains over six million articles, the smaller half of Wikipedias (155 languages) put together only total approximately 400,000 articles. As the BERT base model has over 100 million parameters and BERT models are frequently trained on billions of words of unannotated text, it seems safe to estimate that attempting to train BERT with the data from one of the smaller wikipedias⁷ would likely not produce a very successful model. It is nevertheless not well established how much unannotated text is required to pre-train a language-specific model, and how much the domain and quality of the pre-training data affect the model performance.

In order to focus computational resources on models with practical value, we opted to exclude "dead" languages that are not in everyday spoken use by any community from our efforts. We have

¹https://github.com/wietsedv/bertje

²https://camembert-model.fr/

³https://turkunlp.org/FinBERT/

⁴https://github.com/deepmipt/

⁵The remaining quarter of BERT pre-training data was drawn from the BooksCorpus (Zhu et al., 2015), a unique (and now unavailable) resource for which analogous resources in other languages cannot be readily created.

⁶https://en.wikipedia.org/wiki/List_ of_Wikipedias

⁷For example, Old Church Slavonic, ranked 272nd among wikipedias by size, has fewer than 1000 articles and under 50,000 tokens.

Language (code)	Tokens	Language (code)	Tokens	Language (code)	Tokens
Afrikaans (af)	24M	Finnish (fi)	97M	Norwegian (no)	112M
Arabic (ar)	184M	French (fr)	858M	Polish (pl)	282M
Belarusian (be)	34M	Galician (gl)	58M	Portuguese (pt)	326M
Bulgarian (bg)	71M	Hebrew (he)	166M	Romanian (ro)	85M
Catalan (ca)	236M	Hindi (hi)	35M	Russian (ru)	565M
Czech (cs)	143M	Croatian (hr)	54M	Slovak (sk)	39M
Danish (da)	65M	Hungarian (hu)	129M	Slovenian (sl)	42M
German (de)	1.0B	Indonesian (id)	93M	Serbian (sr)	96M
Greek (el)	81M	Italian (it)	579M	Swedish (sv)	364M
English (en)	2.7B	Japanese (ja)	596M	Tamil (ta)	26M
Spanish (es)	678M	Korean (ko)	79M	Turkish (tr)	71M
Estonian (et)	38M	Lithuanian (lt)	34M	Ukrainian (uk)	260M
Basque (eu)	45M	Latvian (lv)	21M	Urdu (ur)	18M
Persian (fa)	95M	Dutch (nl)	300M	Vietnamese (vi)	172M

Table 1: Wikipedia sizes for selected languages.

otherwise broadly proceeded to introduce preprocessing support and models for languages in decreasing order of the size of their Wikipedias and support in Universal Dependencies, discussed below. Table 1 lists the Wikipedias used in this work.

3.2 Universal Dependencies

Universal Dependencies (UD) is a communitylead effort aiming to create cross-linguistically consistent treebank annotations for many typologically different languages (Nivre et al., 2016, 2020). In this study, we rely on UD both as training data for components of the preprocessing pipeline (Section 4.1) as well as for our evaluations. As of this writing, the latest release of the UD treebanks⁸ is 2.7, which includes 183 treebanks covering 104 languages, thus matching mBERT in terms of the raw number of covered languages.

To maintain comparability with recent work on UD parsing, we use the UD v2.3 treebanks,⁹ with 129 treebanks in 76 languages, in our comparative experiments assessing the WikiBERT models. We further limit our evaluation to the subset of UD v2.3 treebanks that have training, development, and test sets, thus excluding e.g. the 17 parallel UD treebanks which only provide test sets. We further exclude from evaluation treebanks released without text (ar_nyuad, en_esl, fr_ftb, ja_bccwj), the Swedish sign language treebank (swl_sslc), and treebanks in languages for which we have not trained dedicated models (mr_ufal, mt_mudt, te_mtg, and ug_udt). Table 2 lists the treebanks applied in our evaluation. We note that there is very substantial variance between treebanks in the amount of training data available, ranging from little over 3000 tokens for the Lithuanian HSE treebank to more than a million for the Czech PDT.

4 Methods

We next briefly introduce the primary steps of the preprocessing pipeline for creating pre-training examples from Wikipedia source as well as the tools used for text processing, model pre-training, and evaluation. We refer to our published pipeline and its documentation for full processing details.

4.1 Preprocessing pipeline

In order to create high quality pre-training data from raw Wikipedia dumps in the format required by BERT model training, we introduce a pipeline that performs the following primary steps:

Data and model download The full Wikipedia database backup dump is downloaded from a mirror site¹⁰ and a UDPipe model for the language from the LINDAT/CLARIN repository.¹¹

Plain text extraction WikiExtractor¹² is used to extract plain text with document boundaries from the Wikipedia XML dump.

⁸https://universaldependencies.org/

⁹http://hdl.handle.net/11234/1-2895

¹⁰https://dumps.wikimedia.org/

¹¹http://hdl.handle.net/11234/1-3131

¹²https://github.com/attardi/

wikiextractor

Language (code)	Treebank	Tokens	Language (code)	Treebank	Tokens
Afrikaans (af)	AfriBooms	33894	Indonesian (id)	GSD	97531
Arabic (ar)	PADT	223881	Italian (it)	ISDT	276019
Belarusian (be)	HSE	5217	Italian (it)	ParTUT	48934
Bulgarian (bg)	BTB	124336	Italian (it)	PoSTWITA	99441
Catalan (ca)	AnCora	417587	Japanese (ja)	GSD	160419
Czech (cs)	CAC	472609	Korean (ko)	GSD	56687
Czech (cs)	CLTT	26742	Korean (ko)	Kaist	296446
Czech (cs)	FicTree	133637	Lithuanian (lt)	HSE	3210
Czech (cs)	PDT	1173282	Latvian (lv)	LVTB	113405
Danish (da)	DDT	80378	Dutch (nl)	Alpino	186046
German (de)	GSD	263804	Dutch (nl)	LassySmall	75134
Greek (el)	GDT	42326	Norwegian (no)	Bokmaal	243887
English (en)	EWT	204585	Norwegian (no)	Nynorsk	245330
English (en)	GUM	53686	Polish (pl)	LFG	104750
English (en)	LinES	50091	Polish (pl).	SZ	62501
English (en)	ParTUT	43518	Portuguese (pt)	Bosque	206744
Spanish (es)	AnCora	444617	Portuguese (pt)	GSD	255755
Spanish (es)	GSD	382436	Romanian (ro)	Nonstandard	155498
Estonian (et)	EDT	341122	Romanian (ro)	RRT	185113
Basque (eu)	BDT	72974	Russian (ru)	GSD	75964
Persian (fa)	Seraji	121064	Russian (ru)	SynTagRus	870474
Finnish (fi)	FTB	127602	Slovak (sk)	SNK	80575
Finnish (fi)	TDT	162621	Slovenian (sl)	SSJ	112530
French (fr)	GSD	354699	Serbian (sr)	SET	65764
French (fr)	ParTUT	24123	Swedish (sv)	LinES	48320
French (fr)	Sequoia	50536	Swedish (sv)	Talbanken	66645
French (fr)	Spoken	14952	Tamil (ta)	TTB	6329
Galician (gl)	CTG	79327	Turkish (tr)	IMST	37918
Hebrew (he)	HTB	137721	Ukrainian (uk)	IU	88043
Hindi (hi)	HDTB	281057	Urdu (ur)	UDTB	108690
Croatian (hr)	SET	154055	Vietnamese (vi)	VTB	20285
Hungarian (hu)	Szeged	20166			•

Table 2: UD v2.3 training data sizes for selected treebanks.

Segmentation and tokenization UDPipe is used with the downloaded model to segment sentences and tokenize the plain text, producing text with document, sentence, and word boundaries.

Document filtering A set of heuristic rules and statistical language detection¹³ are applied to optionally filter documents based on configurable criteria.¹⁴

Sampling and basic tokenization A sample of sentences is tokenized using BERT basic tokeniza-

tion¹⁵ to produce examples for vocabulary generation that match BERT tokenization criteria.

Vocabulary generation A subword vocabulary is generated using the SentencePiece¹⁶ (Kudo and Richardson, 2018) implementation of byte-pair encoding (Gage, 1994; Sennrich et al., 2015). After generation the vocabulary is converted to the BERT WordPiece format (a different but largely equivalent representation).

¹³https://github.com/shuyo/

language-detection

¹⁴We note that there are Wikipedia pages whose content is mostly in a language different from that of the Wikipedia.

¹⁵BERT basic tokenization preserves alphanumeric sequences but separates e.g. all punctuation characters into individual tokens.

¹⁶https://github.com/google/ sentencepiece

	Subword Accuracy		Subword Accuracy		
Language (code)	mBERT	WikiBERT	Language (code)	mBERT	WikiBERT
Afrikaans (af)	28.69	43.22	Indonesian (id)	30.72	52.47
Arabic (ar)	20.17	29.96	Italian (it)	29.48	37.98
Belarusian (be)	18.15	36.39	Japanese (ja)	49.25	45.19
Bulgarian (bg)	21.26	39.98	Korean (ko)	17.59	30.61
Catalan (ca)	40.29	56.63	Lithuanian (lt)	15.11	29.83
Czech (cs)	22.41	39.77	Latvian (lv)	15.59	29.99
Danish (da)	25.06	40.86	Dutch (nl)	29.08	47.54
German (de)	33.85	46.93	Norwegian (no)	22.73	34.15
Greek (el)	21.42	45.42	Polish (pl)	17.64	33.30
English (en)	37.39	46.64	Portuguese (pt)	32.55	43.85
Spanish (es)	40.20	52.05	Romanian (ro)	21.19	33.07
Estonian (et)	14.00	31.26	Russian (ru)	27.16	46.86
Basque (eu)	15.15	30.99	Slovak (sk)	16.52	29.08
Persian (fa)	21.52	45.20	Slovenian (sl)	21.21	35.24
Finnish (fi)	12.89	27.67	Serbian (sr)	25.80	30.70
French (fr)	41.30	52.08	Swedish (sv)	22.11	37.11
Galician (gl)	33.23	36.81	Tamil (ta)	14.36	31.85
Hebrew (he)	20.96	21.83	Turkish (tr)	12.56	29.16
Hindi (hi)	19.97	47.23	Ukrainian (uk)	19.15	31.78
Croatian (hr)	23.03	39.99	Urdu (ur)	20.83	39.70
Hungarian (hu)	18.89	38.99	Vietnamese (vi)	17.96	47.35

Table 3: Results for the cloze test in terms of subword prediction accuracy (percentages)

Example generation Masked language modeling and next sentence prediction examples using the full BERT tokenization specified by the generated vocabulary are created in the TensorFlow TFRecord format using BERT tools.

The created vocabulary and pre-training examples can be used directly with the original BERT implementation to train new language-specific models.

4.2 UDPipe

UDPipe (Straka et al., 2016) is a parser capable of producing segmentation, part-of-speech and morphological tags, lemmas and dependency trees. In this work we use UDPipe for sentence segmentation and tokenization in the preprocessing pipeline. The segmentation component in UDPipe is a character-level bidirectional GRU network simultaneously predicting the end-of-token and endof-sentence markers.

4.3 Pre-training

We aimed to largely mirror the original BERT process in our selection of parameters and settings for the pre-training process to create the Wiki-BERT models, with some adjustments made to account for differences in computational resources. Specifically, while the original BERT models were trained on TPUs, we trained on Nvidia Volta V100 GPUs with 32GB memory. We followed the original BERT processing in training for a total of 1M steps in two stages, the first 900K steps with a maximum sequence length of 128, and the last 100K steps with a maximum of 512. Due to memory limitations, each model was trained on 4 GPUs using a batch size of 140 during the sequence length 128 phase, and 8 GPUs with a batch size of 20 during the sequence length 512 phase.

4.4 Cloze test

In order to evaluate the BERT models with respect to their original training objective, we employ a cloze test, where words are randomly masked and predicted back. We mask a random 15% of words in each sentence, and, in case a word is composed of several subword (WordPiece) tokens, all subword tokens are masked for an easier and more meaningful evaluation (cf. full-word masking in BERT pre-training). All masked positions are predicted at once in the same manner as done in the BERT pre-training (i.e. without iterative predic-

	Average LAS		Average LAS		
Language (code)	mBERT	WikiBERT	Language (code)	mBERT	WikiBERT
Afrikaans (af)	87.85	87.33	Indonesian (id)	80.40	80.12
Arabic (ar)	83.81	85.47	Italian (it)	89.64	89.77
Belarusian (be)	81.77	79.81	Japanese (ja)	92.78	92.92
Bulgarian (bg)	92.30	92.51	Korean (ko)	86.19	87.28
Catalan (ca)	92.08	92.06	Lithuanian (lt)	58.68	58.40
Czech (cs)	90.45	90.69	Latvian (lv)	84.29	84.46
Danish (da)	85.78	85.84	Dutch (nl)	90.26	91.02
German (de)	83.16	84.13	Norwegian (no)	91.54	91.94
Greek (el)	91.63	92.35	Polish (pl)	94.45	95.58
English (en)	88.09	88.05	Portuguese (pt)	91.91	92.21
Spanish (es)	90.42	90.12	Romanian (ro)	86.83	86.52
Estonian (et)	85.86	87.43	Russian (ru)	90.35	91.13
Basque (eu)	82.99	83.70	Slovak (sk)	91.64	91.73
Persian (fa)	86.60	88.60	Slovenian (sl)	92.83	93.37
Finnish (fi)	87.64	90.81	Serbian (sr)	92.30	91.79
French (fr)	89.22	88.77	Swedish (sv)	86.42	87.12
Galician (gl)	83.05	82.61	Tamil (ta)	70.14	69.63
Hebrew (he)	88.77	90.17	Turkish (tr)	69.33	71.25
Hindi (hi)	91.59	91.86	Ukrainian (uk)	88.57	90.41
Croatian (hr)	89.46	89.40	Urdu (ur)	82.66	82.15
Hungarian (hu)	83.99	86.21	Vietnamese (vi)	66.89	68.87

Table 4: Average LAS results for UDify for Universal Dependencies treebanks in each language.

tion of one position per time step). As a source of sentences, we use the first 1000 sentences of training sections of the treebanks, limited to sentences of 5–50 tokens in length. We note that the treebanks are not entirely non-overlapping with Wikipedia: 16 out of the 63 treebanks draw at least part of their texts from Wikipedia. However, as all of the compared models share this source of pretraining data, we do not expect this overlap to bias the comparison.

4.5 UDify

To assess the performance of the models in a downstream task, we apply the UDify parser (Kondratyuk and Straka, 2019), initialized with one of the models and trained on Universal Dependencies data. UDify is a state-of-the-art model and can predict UD part-of-speech tags, morphological features, lemmas, and dependency trees. UDify implements a multi-task learning objective using task-specific prediction layers on top of a pre-trained BERT encoder. All prediction layers are trained simultaneously, while also finetuning the pre-trained encoder weights. In the following evaluation, we focus on the parsing performance using the standard Labeled Attachment Score (LAS) metric.

5 Results

We next present the results of the intrinsic cloze test evaluation and the extrinsic evaluation with syntactic analysis as a downstream task.

5.1 Cloze evaluation results

The cloze evaluation results are shown in Table 3, where we measure subword-level prediction accuracy, i.e. the proportion of cases where the model assigns the highest probability to the original subword. We find that the WikiBERT models outperform mBERT for all languages except for Japanese,¹⁷ averaging more than 10% points higher accuracy. While this is an encouraging result regarding the quality of the newly introduced models, the evaluation is arguably biased in favour of monolingual models, as their candidate space (the vocabulary) is limited to only include options in the correct language. More broadly, success at

¹⁷This result may suggest some issues specific to Japanese either in the preprocessing pipeline or the applied UDify model, but we have yet to identify any clear explanation for the exception.



Figure 1: Average relative change in LAS when replacing mBERT with a WikiBERT model for UDify initialization plotted against the WikiBERT pre-training data size in tokens. Coloring indicates language grouping by genera (Baltic: white, Finnic: light blue, Germanic: yellow, Indic: orange, Romance: red, Semitic: green, Slavic: blue, other: black).

intrinsic evaluations such as this does not guarantee practical applicability (or vice versa), and models should also be assessed at real-world tasks to gain a more complete picture of their value (see e.g. Chiu et al., 2016).

5.2 UD parsing results

Table 4 summarizes the results of the UD parsing evaluation. Given the large size of both train sets (See Table 2) and test sets for most of the languages, the evaluation results are stable, and we have found that repetitions of the training process often result in less than 0.1% point differences between runs. To conserve computational resources, we have thus here chosen to run a single experiment per treebank (a typical setting for UD evaluation).

We find a complex, mixed picture where mBERT and WikiBERT models each appear clearly superior for different languages, for example, mBERT for Belarusian and WikiBERT for Finnish. On average across all languages, UDify with WikiBERT models slightly edges out UDify with mBERT, with an 86.1% average for mBERT and 86.6% for WikiBERT (an approximately 4%) relative decrease in LAS error). However, such averaging hides more than it reveals, and it is more interesting to consider the various potential impacts on performance from pre-training data size, potential support from close relatives in the same language family, and other similar factors. The various UD treebanks represent very different levels of challenge with LAS results ranging from below 60% to above 95%, and to reduce the impact of the properties of the treebanks on the comparison, in the following we focus on the relative change in performance when initializing UDify with a WikiBERT model compared to the baseline approach using mBERT.

Figure 1 shows the average relative change in performance over all treebanks for a language when replacing mBERT with the relevant Wiki-BERT model for UDify, plotted against the number of tokens in Wikipedia for the language. While the data is very noisy due to a number of factors, we find some indication of a "sweet spot" where training a dedicated monolingual model tends to show most benefit over using the multilingual model when at least approximately 100M tokens but fewer than 1B tokens of pre-training data are available. We also briefly note some other properties in this data:

- For English, a language in the large Germanic family and the one with the largest amount of Wikipedia pre-training data, mBERT and WikiBERT results are effectively identical.
- The greatest loss when moving from mBERT to a WikiBERT model is seen for Belarusian, a slavic language closely related to Russian, for which considerably more pretraining data is available.
- The greatest gain when moving from mBERT to a WikiBERT model is seen for Finnish, a Finnic language with few closely related, widely spoken languages, which has a comparatively large Wikipedia.

Observations such as these may suggest fruitful avenues for further research into the conditions under which mono- and multilingual language model training is expected to be most successful. Based on these results and the findings of studies training models for small numbers of closely related languages (see Section 2), we anticipate that multilingual training may most readily benefit lower-resourced languages trained together with a closely related high-resource language in a bilingual setting.

6 Discussion and conclusions

In this paper, we have introduced a simple, fully automatic pipeline for creating monolingual BERT models from Wikipedia data, applied the pipeline to introduce 42 new language-specific models, most covering languages that previously lacked a dedicated deep neural language model. We evaluated the WikiBERT models intrinsically using cloze evaluation, finding that they outperform the multilingual mBERT model for all but one language. An extrinsic evaluation using a dependency parsing task with Universal Dependencies data and the UDify neural parser found a more nuanced picture of the comparative merits of the monolingual and multilingual models: while we found that a WikiBERT model will provide better performance than mBERT on average and in multiple cases provides a more than 10% relative decrease in LAS error compared to the multilingual model, the WikiBERT models showed lower performance than mBERT for multiple languages. Viewing relative change in performance against pre-training data size, we found indications that monolingual models may most benefit languages that have no closely related highresource languages and for which comparatively large pre-training corpora can be assembled.

The availability of the WikiBERT collection of models opens up a broad range of potential avenues for research into the strengths, weaknesses and challenges in both mono- and multilingual language modeling that we hope to pursue in future work. We also hope to encourage both monolingual applications as well as exploration of these questions by others by making the models freely available under open licenses from https:// github.com/turkunlp/wikibert.

Acknowledgments

This work was funded in part by the Academy of Finland. We wish to thank CSC – IT Center for Science, Finland, for providing generous computational resources for this study.

References

- Li-Hsin Chang, Sampo Pyysalo, Jenna Kanerva, and Filip Ginter. 2020. Towards fully bilingual deep language modeling. *arXiv preprint arXiv:2010.11639*.
- Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440– 8451.
- 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.

- Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of Romanian BERT. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4324–4328.
- Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- Pentii Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 22.
- Daniel Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.
- Taku Kudo and John Richardson. 2018. Sentence-Piece: 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.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
- 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. *arXiv preprint arXiv:1907.11692*.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty french language model. In *Proceedings of the 58th Annual Meeting* of the Association for Computational Linguistics.
- Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations* (*ICLR 2013*), pages 1–12.
- Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1659–1666.

- 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. arXiv preprint arXiv:2004.10643.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? *arXiv* preprint arXiv:1906.01502.
- Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.
- Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290– 4297.
- Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331.
- Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. *arXiv preprint arXiv:2006.07890*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for finnish. *arXiv preprint arXiv:1912.07076*.

- Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A dutch BERT model. *arXiv preprint arXiv:1912.09582*.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3261–3275.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision* (*ICCV*), pages 19–27.