Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2050–2057
Marseille, 20-25 June 2022

EENLP: Cross-lingual Eastern European NLP Index

Alexey Tikhonov*, **Alex Malkhasov**, **Andrey Manoshin**, **George Dima**,
Réka Cserháti, **Md.Sadek Hossain Asif**, **Matt Sárdi**

Independent researcher, Germany; Financial University of Russia; MEPhI, Russia;
University Politehnica of Bucharest, Romania; University of Szeged, Hungary;
Notre Dame College, Dhaka, Bangladesh; Mozaik Education, Hungary

{altsoph, realex1902, sqidde, andreigeorgedima, cserhatir, asifsadek509, sardi.matt}@gmail.com

Abstract

Motivated by the sparsity of NLP resources for Eastern European languages, we present a broad index of existing Eastern European language resources (90+ datasets and 45+ models) published as a github repository open for updates from the community. Furthermore, to support the evaluation of commonsense reasoning tasks, we provide hand-crafted cross-lingual datasets for five different semantic tasks (namely news categorization, paraphrase detection, Natural Language Inference (NLI) task, tweet sentiment detection, and news sentiment detection) for some of the Eastern European languages. We perform several experiments with the existing multilingual models on these datasets to define the performance baselines and compare them to the existing results for other languages.

Keywords: cross-lingual, less-resourced languages, NLP

1. Introduction

Recent multilingual Transformer-based language models – such as mBERT (Devlin et al., 2018), XLM-RoBERTa (Sanh et al., 2019), multilingual DistilBERT (Conneau et al., 2020), etc. – show impressive results on different text analysis tasks and their cross-lingual reasoning capabilities are still actively studied (Lauscher et al., 2020). For example, the mBERT model trained on 104 languages has shown high cross-lingual performance; however, such evaluations mostly focused on cross-lingual transfer within high-resource languages (Wu and Dredze, 2020).

Commonsense reasoning is one of the key problems in natural language processing, but the relative scarcity of labeled data holds back the progress for languages other than English: there are widely spoken languages that still did not receive the focus of the research community (Joshi et al., 2020). Cross-lingual transfer learning could be beneficial for such languages in solving both theoretical and practical tasks. One can speculate that such high-level reasoning tasks could be less affected by the language syntax's specifics and low-level properties, so it can be effective to use a cross-lingual approach here. However, it was shown that the success of cross-lingual transfer learning depends on different factors such as the amount of shared vocabulary, explicit alignment of representations across languages, size of pretraining corpora, etc (Doddapaneni et al., 2021). To get a better understanding of the importance of these factors, researchers need to leverage diverse and detailed datasets.

There are a bunch of cross-lingual datasets already, such as XGLUE (Liang et al., 2020), XCOPA (Ponti et al., 2020), XL-WiC (Raganato et al., 2020), XWINO (Tikhonov and Ryabinin, 2021), Wino-X (Emelin and Sennrich, 2021), XNLI (Conneau et al., 2018), XL-WSD (Pasini et al., 2021), XTREME-R (Ruder et al., 2021), BSNLP (Piskorski et al., 2019), MOROCO (Butnaru and Ionescu, 2019), etc., but most of them cover high-resource languages.

In this paper, we concentrate on Eastern European languages. These languages are numerous and heterogeneous; they include languages from at least two language families (Indo-European and Uralic), and the former family is represented by very diverse branches. Although there are several dedicated multi-task benchmarks for a few Eastern European languages – e.g., KLEJ (Rybak et al., 2020) for Polish, LiRo (Dumitrescu et al., 2021) for Romanian, RussianSuperGLUE (Shavrina et al., 2020) for Russian, or the translation of SuperGLUE for Slovene, they usually concentrate on one or two languages. They also use different tasks and various data formats; hence they can not be used for cross-lingual benchmarking without careful manual pre-processing.

The main contributions of this paper are:

- To build a comprehensive picture of the current NLP state for the Eastern European languages, we present the wide index of existing Eastern European languages resources (90+ datasets and 45+

* The corresponding author is Alexey Tikhonov

models) published as a github repository[1] open for updates from the community;

- Next, we provide hand-crafted cross-lingual datasets for five different semantic tasks (namely news categorization, paraphrase detection, Natural Language Inference (NLI) task, tweet sentiment detection, and news sentiment detection), compiled by processing data from various sources into the same format, opening room for evaluation scenarios such as zero-shot cross-lingual transfer. Since the source datasets are licensed under various licenses, we publish automatic scripts for our datasets compilation on the same github repository;
- Finally, we perform several experiments with the existing multilingual models on our datasets to define the performance baselines and compared them to the existing similar results for other languages.

We made all our code and data publicly available. We also published the detailed results of our experiments at our Weights and Biases project[2] (Biewald, 2020).

2. Dataset index

Aiming to build the index of Eastern European language NLP resources useful to the community, we have created a long list of available datasets we were able to find. We focused on supervised datasets (with data labels) and preferred semantic tasks to syntactic ones. Lastly, since some languages are already well covered with benchmarks (for example, Polish, Romanian, Russian), we tried to focus on less represented languages. Indeed, we do not claim this list is exhaustive; however, we provide an easy way to add any missing resources via the creation of a github issue. We encourage the community to help with updating this index further.

For this moment, we've already collected more than 90 datasets for 20 different languages: Albanian, Armenian, Belarusian, Bosnian, Bulgarian, Croatian, Czech, Estonian, Hungarian, Latvian, Lithuanian, Macedonian, Moldavian, Polish, Romanian, Russian, Serbian, Slovak, Slovenian, and Ukrainian.

These datasets cover various tasks, including text category prediction, coreference resolution, fake news detection, lemmatization, morphosyntactic tagging, NER, NLI, offensive comments detection, paraphrase detection, POS tagging, question answering, sentiment analysis, word sense disambiguation, and many more. The complete list of the discovered datasets is published on our github repository[3].

3. Models index

In a similar way, we tried to enumerate available models useful for Eastern European languages. We were aiming for the Transformer-based MLM models (which are state of the art for text classification nowadays), but we also found some models of other types, including Causal Language Models, Adapters, static embeddings, etc; we listed them as well.

We started with the modern well-known multilingual models, namely mBERT (Devlin et al., 2018), mDistilBERT (Sanh et al., 2019), XLM-Roberta (Conneau et al., 2020), and LaBSE (Reimers and Gurevych, 2019). Next, there are several cross-lingual models, such as CroSloEngual (Ulcar and Robnik-Sikonja, 2020), BERTić (Ljubesic and Lauc, 2021), SlavicBERT (Arkhipov et al., 2019), or LitLat BERT. Finally, we found more than 20 language-specific BERT-like models; most of them are listed in Figure 1.

Besides MLM models, we found 6 GPT-like models and several dozens of models from the pre-Transformer era: ULMFiT, ELMo, and static word embeddings. All our findings are available on our github repository[4] as well.

4. Benchmark tasks

Considering the coverage and sparsity of source datasets, we decided to proceed with five commonsense reasoning tasks with different Eastern European languages subsets. We checked these tasks for the languages coverage and finally decided to focus on this short-list:

- news categorisation,
- paraphrase detection,
- news sentiment detection,
- tweet sentiment detection,
- NLI.

For each of these tasks, we manually crafted a dedicated dataset with maximized Eastern European languages coverage. Whenever it was possible we also added English as a language most common for contemporary pre-trained models. The properties of these datasets are given in the following sections. Since the source datasets are licensed under various licenses, we published automatic scripts for our datasets compilation on the same github repository[5].

For each of these tasks, we evaluated three modern transformer-based pre-trained models: a multilanguage version of BERT, a model XLM-Roberta that is two times larger than BERT, and a distilled version of multilanguage BERT (to check how the distillation

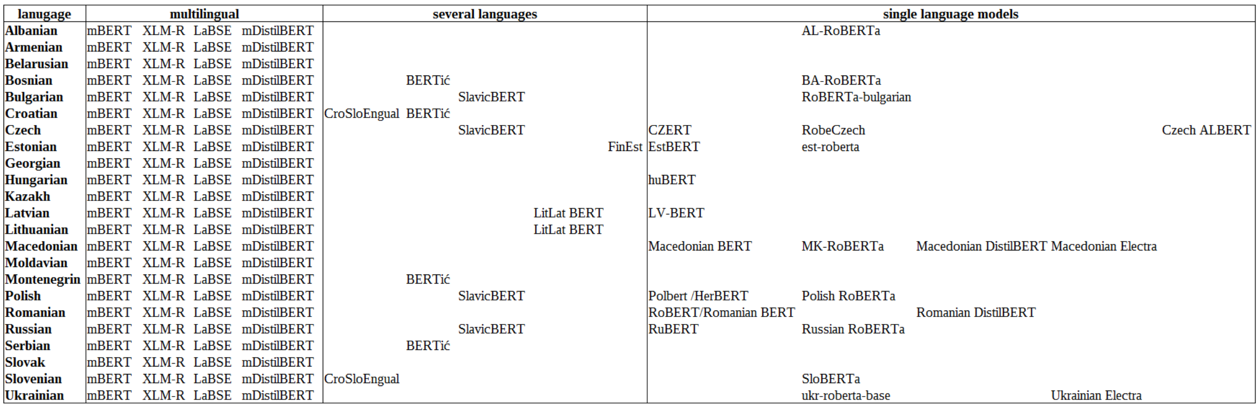
[1] https://github.com/altsoph/EENLP

[2] https://wandb.ai/eenlp

[3] https://github.com/altsoph/EENLP/blob/main/docs/datasets.md

[4] https://github.com/altsoph/EENLP/blob/main/docs/models.md

[5] https://github.com/altsoph/EENLP/tree/main/build_benchmarks



lanugage	multilingual				several languages					single language models				
Albanian	mBERT	XLM-R	LaBSE	mDistilBERT							AL-RoBERTa			
Armenian	mBERT	XLM-R	LaBSE	mDistilBERT										
Belarusian	mBERT	XLM-R	LaBSE	mDistilBERT										
Bosnian	mBERT	XLM-R	LaBSE	mDistilBERT		BERTić					BA-RoBERTa			
Bulgarian	mBERT	XLM-R	LaBSE	mDistilBERT			SlavicBERT				RoBERTa-bulgarian			
Croatian	mBERT	XLM-R	LaBSE	mDistilBERT	CroSloEngual	BERTić								
Czech	mBERT	XLM-R	LaBSE	mDistilBERT			SlavicBERT			CZERT	RobeCzech			Czech ALBERT
Estonian	mBERT	XLM-R	LaBSE	mDistilBERT					FinEst	EstBERT	est-roberta			
Georgian	mBERT	XLM-R	LaBSE	mDistilBERT										
Hungarian	mBERT	XLM-R	LaBSE	mDistilBERT						huBERT				
Kazakh	mBERT	XLM-R	LaBSE	mDistilBERT										
Latvian	mBERT	XLM-R	LaBSE	mDistilBERT				LitLat BERT		LV-BERT				
Lithuanian	mBERT	XLM-R	LaBSE	mDistilBERT				LitLat BERT						
Macedonian	mBERT	XLM-R	LaBSE	mDistilBERT						Macedonian BERT	MK-RoBERTa	Macedonian DistilBERT	Macedonian Electra	
Moldavian	mBERT	XLM-R	LaBSE	mDistilBERT										
Montenegrin	mBERT	XLM-R	LaBSE	mDistilBERT		BERTić								
Polish	mBERT	XLM-R	LaBSE	mDistilBERT			SlavicBERT			Polbert /HerBERT	Polish RoBERTa			
Romanian	mBERT	XLM-R	LaBSE	mDistilBERT						RoBERT/Romanian BERT		Romanian DistilBERT		
Russian	mBERT	XLM-R	LaBSE	mDistilBERT			SlavicBERT			RuBERT	Russian RoBERTa			
Serbian	mBERT	XLM-R	LaBSE	mDistilBERT		BERTić								
Slovak	mBERT	XLM-R	LaBSE	mDistilBERT										
Slovenian	mBERT	XLM-R	LaBSE	mDistilBERT	CroSloEngual						SloBERTa			
Ukrainian	mBERT	XLM-R	LaBSE	mDistilBERT							ukr-roberta-base		Ukrainian Electra	

Figure 1: MLM Transformer-based models with Eastern European languages coverage.

affects the cross-lingual transfer quality). Whenever the English language was also available for the task we added a base English BERT model to our evaluation. We fine-tuned them using AdamW optimizer with the learning rate 1e-5, epsilon 1e-8, and the linear schedule with warmup. We reported all the metrics for the epoch with the best source language validation result and published our code for evaluation[6].

We will now describe every task in detail.

4.1. News categorisation task

In this task, the model should detect the correct category of the news text. This is a basic semantic task; however, it's not always obvious how to separate, for example, categories like ”lifestyle” and ”entertainment.” Our dataset consists of 8 major categories and covers 7 Eastern European languages + English (however, the distribution of categories differs from one language to another, consider Table 1).

The sources for this dataset are:

- English: (Liang et al., 2020),
- Armenian: (Avetisyan and Ghukasyan, 2019),
- Estonian: (Purver et al., 2021),
- Latvian: (Pollak et al., 2021),
- Moldavian: (Butnaru and Ionescu, 2019),
- Romanian: (Butnaru and Ionescu, 2019),
- Russian: (Liang et al., 2020),
- Slovak: (Hladek et al., 2014).

We use this task to measure the models' cross-lingual transfer quality while transferring from English to various Eastern European languages. (One can also use Latvian or Russian as a source language since both of them cover all of the categories). Following (Liang et al., 2020), we use multi-class classification accuracy as the key metric for this task. Table 2 sums up the results of four models on a multilanguage dataset. Further details across languages and categories are available on W&B[7].

The main observations based on the results of this task are:

- multilingual pre-training is the key factor for successful cross-lingual transfer learning, English BERT model shows no ability to transfer its knowledge from English to other languages;
- mBERT and XLM-R models show a similar level of quality as in (Liang et al., 2020) for the similar NC task (however, note, it's unfair to compare them directly since the list of categories is not the same);
- distillation significantly affects the ability to cross-lingual transfer learning; this effect is significant for almost all of the tested languages;
- the results vary across the languages a lot, this is a direction of possible deeper analysis (note that, for example, Romanian and Moldavian are close languages; however, all the models are performing significantly better on Moldavian);
- mBERT generally performs slightly better than XLM-RoBERTa, this could be an effect of relatively small datasets.

4.2. Paraphrase detection task

In the paraphrase detection task, the model should decide whether a pair of sentences have the same meaning. It is another standard task for contemporary NLU benchmarks; however, it is significantly hard and can be problematic even for the English language (Yamshchikov et al., 2021). Our dataset consists of four Eastern European languages + English. The sources for this dataset are:

- English: (Wang et al., 2018),

[6] https://github.com/altsoph/EENLP/tree/main/eval_benchmarks

[7] https://wandb.ai/eenlp/newscat_full

	en	**hy**	**ee**	**lv**	**mo**	**ro**	**ru**	**sk**
entertainment	3498	996	275	227	1234	1058	1498	22
finance	13405	2066	13	34	3429	5105	2514	1022
health	5751		121	119			311	559
lifestyle	6295	1630	107	37			671	481
news/accidents	31490	1454	33526	3722			12125	1655
sports	38598	2797		149	3443	2583	1731	1042
tech/auto	3926		144	40	1266	3392	387	
travel	3076		49	55			53	

Table 1: News categorisation dataset properties. For each language / category pair we reported the number of the items (short news texts) in our dataset.

	mDBERT	mBERT	XLMR	BERT
en	91.11%	**92.50**%	92.36%	92.65%
ee	74.83%	**81.58**%	72.55%	27.33%
hy	55.57%	66.10%	**67.64%**	21.41%
lv	63.79%	86.60%	**87.68%**	6.31%
mo	65.79%	**66.11**%	62.36%	16.12%
ro	46.18%	**47.54**%	44.18%	7.46%
ru	77.49%	**77.34**%	77.10%	43.31%
sk	59.88%	**63.25**%	60.32%	35.34%

Table 2: News categorisation task. For each language / model pair we reported the accuracy value on the same epoch where the validation accuracy was the best. We also evaluated a standard English BERT here to compare.

Language	Items	Classes
English	373 263	2
Armenian	4 233	2
Polish	8 000	2
Romanian	5 749	2
Serbian	835	2

Table 3: Paraphrase detection dataset properties. Each item is a pair of sentences.

- Armenian: (Malajyan et al., 2020),
- Polish: (Rybak et al., 2020),
- Romanian: RO-STS dataset[8],
- Serbian: (Batanović et al., 2011).

We also tried to leverage the TaPaCo dataset (Scherrer, 2020) to build a derivative dataset with positive and negative examples of paraphrases. We treated pairs of sentences from the same cluster as positive paraphrase examples. Then, we used several strategies for sampling hard-negative samples (including sampling from similar sentences by Levenshtein distance and LabSE (Feng et al., 2020) embedding distances). To assess the quality of such dataset augmentation, we performed a two-folded cross-check: considering languages where we have both TaPaCo and non-TaPaCo datasets, we checked the cross-datasets transfer learning in both directions. The results of non-TaPaCo->TaPaCo transfer were much better than for the opposite direction (for example, for Polish, it was 68% versus 51%), which implies that datasets represent different tasks with non-equal complexity. Thus, we decided to reject a TaPaCo-based augmentation until further investigations. However, we consider it an interesting direction for future work.

	mDBERT	mBERT	XLMR	BERT
en	81.84%	82.20%	**83.80**%	82.20%
hy	88.68%	**92.79**%	92.70%	75.36%
pl	87.86%	**87.29**%	82.14%	84.33%
ro	79.82%	82.36%	**86.88**%	80.03%
sr	72.10%	**77.01**%	71.74%	55.09%

Table 4: Paraphrase detection task. For each language / model pair we reported the accuracy value on the same epoch where the validation accuracy was the best. We also evaluated a standard English BERT here to compare.

We used the same setup as in the previous task – cross-lingual zero-shot learning from English to other languages. Following (Yang et al., 2019), we use classification accuracy as the key metric for this task. We reported values in table 4 from the same epoch where the validation accuracy was the best. More details are available online[9].

The main observations based on these results are:

- again, multilingual pre-trained models show consistently better performance than English BERT, which highlights the importance of the multilingual pre-training phase;
- mBERT results are competitive with the results reported in (Yang et al., 2019);
- once more, distillation affects the ability to cross-lingual transfer learning.

[8] https://github.com/dumitrescustefan/RO-STS

[9] https://wandb.ai/eenlp/paraphrase_detection/artifacts/paraphrase_detection-dataset/paraphrase_detection-dataset/d144e76e6129b022dc5d

4.3. News sentiment detection task

Language	Items	Classes
Croatian	1 500	3
Lithuanian	10 000	3
Russian	8 200	3
Slovenian	10 000	3

Table 5: News sentiment detection dataset properties. Each item is a short news text.

In this task, we made an evaluation of sentiment detection (3 classes: positive, neutral, negative) for relatively long texts (news articles). We compiled a dataset, that covers 4 Eastern European languages (Croatian, Lithuanian, Russian, and Slovenian). Table 5 describes volumes of data per language. The sources for this dataset are:

- Croatian: (Pelicon et al., 2020),
- Lithuanian: TN2gramos [10],
- Russian: Sentiment Analysis in Russian[11],
- Slovenian: (Bučar et al., 2018).

We used classification accuracy as our main metric and evaluated all possible directions of transfer learning for three multilingual models (mBERT, mDBERT, and XLM-R), as in (Pelicon et al., 2020). Check table 6 for the results, more details are available on W&B[12].

This task results support several observations:

- as in previous experiments, the distillation decreases the ability to cross-lingual transfer learning;
- the cross-lingual transfer quality seems to be correlated with the degree of language affinity;
- XLM-RoBERTa generally dominates over the rest of the models.

		hr	lt	ru	sl
hr	mDBERT	67%	20%	48%	53%
	mBERT	66%	23%	53%	57%
	XLMR	**70**%	**24**%	**50**%	**60**%
lt	mDBERT	45%	70%	50%	51%
	mBERT	49%	71%	57%	49%
	XLMR	**51**%	**78**%	**58**%	**63**%
ru	mDBERT	**63**%	39%	68%	52%
	mBERT	57%	**39**%	**71**%	55%
	XLMR	56%	65%	70%	**59**%
sl	mDBERT	**55**%	41%	57%	64%
	mBERT	54%	52%	58%	67%
	XLMR	51%	**59**%	**60**%	**69**%

Table 6: News sentiment detection task. Validation accuracy.

4.4. Twitter sentiment detection task

Language	Items	Classes
Czech	10 000	3
Latvian	1 160	3
Russian	230 000	2
Slovak	1 600	3

Table 7: Twitter sentiment detection dataset properties. Each item is a tweet text.

In addition to the previous task, we considered the task of short texts (tweets) sentiment detection (same 3 classes: positive, neutral, negative). The dataset for this task also covers 4 Eastern European languages (Czech, Latvian, Russian, and Slovak). Table 7 describes volumes of data per language. The sources for this dataset are:

- Czech: (Habernal et al., 2013),
- Latvian: latvian-tweet-sentiment-corpus [13],
- Russian: mokoron[14],
- Slovak: Sentigrade[15].

As previously, we used classification accuracy as our main metric and evaluated all possible directions of transfer learning for three multilingual models (mBERT, mDBERT, and XLM-R). Check table 8 for the results, more details are available on W&B[16]

This task results support several observations:

- again, the distilled model shows the worst ability to cross-lingual transfer learning;

		ru	sk	lv	cz
ru	mDBERT	88%	61%	75%	68%
	mBERT	89%	60%	76%	67%
	XLMR	**95**%	**72**%	**78**%	**79**%
sk	mDBERT	72%	51%	30%	35%
	mBERT	**76**%	59%	32%	38%
	XLMR	72%	**70**%	**38**%	**42**%
lv	mDBERT	39%	35%	70%	52%
	mBERT	**56**%	**38**%	71%	51%
	XLMR	32%	**38**%	**73**%	**57**%
cz	mDBERT	36%	47%	60%	74%
	mBERT	**44**%	50%	60%	76%
	XLMR	**44**%	**62**%	**71**%	**79**%

Table 8: Twitter sentiment detection task. Validation accuracy.

[10] https://www.kaggle.com/rokastrimaitis/lithuanian-financial-news-dataset-and-bigrams

[11] https://www.kaggle.com/c/sentiment-analysis-in-russian

[12] https://wandb.ai/eenlp/news_sentiment

[13] https://github.com/FnTm/latvian-tweet-sentiment-corpus

[14] http://study.mokoron.com/

[15] https://sentigrade.fiit.stuba.sk/data

[16] https://wandb.ai/eenlp/twit_sentiment

- again, the similarity of the languages matters for the cross-lingual transfer quality;
- again, XLM-RoBERTa generally dominates over the rest of the models regardless of the transfer pair.

4.5. NLI task

Language	Items	Classes
English	30 000	3
Bulgarian	30 000	3
Polish	9 000	3
Russian	30 000	3
Slovenian	306	3

Table 9: NLI dataset properties. Each item is a pair of sentences.

Finally, we proposed the evaluation on the natural language inference (NLI) task. NLI is the task of predicting whether a hypothesis sentence is true (entailment), false (contradiction), or undetermined (neutral) given a premise sentence. Our dataset for this task also covers 4 Eastern European languages (Bulgarian, Polish, Russian, and Slovenian) + English. Table 9 describes volumes of data per language. The sources for this dataset are:

- English: (Liang et al., 2020),
- Bulgarian: (Liang et al., 2020),
- Polish: (Rybak et al., 2020),
- Russian: (Liang et al., 2020),
- Slovenian: (Žagar et al., 2020).

As in (Artetxe and Schwenk, 2019) and (Nooralahzadeh et al., 2020), we used classification accuracy as the metric for the NLI task and evaluated all possible directions of transfer learning for the same three multilingual models (mBERT, mDBERT, and XLM-R). Table 10 provides the results, more details are available on W&B[17]

This task results support several observations:

- this task seems to be significantly harder than others;
- there is no consistent winner among the models for this task.

5. Conclusions and discussion

In this project, we have made and published a broad index of NLP resources for Eastern European languages, which, we hope, will be helpful for the NLP community. We have invested in the creation of new cross-lingual datasets focused on Eastern European languages, hand-crafted the benchmarks for 5 common reasoning tasks, and provided the evaluations of several modern multilingual models. We have published all our code to support future research.

		en	**pl**	**ru**	**bg**	**sl**
	mDBERT	65%	**62**%	56%	55%	34%
en	mBERT	69%	48%	59%	58%	**37**%
	XLMR	**76**%	48%	**69**%	**72**%	36%
	mDBERT	40%	92%	**39**%	**40**%	**26**%
pl	mBERT	42%	92%	38%	38%	19%
	XLMR	**46**%	**94**%	38%	**40**%	14%
	mDBERT	62%	**57**%	54%	55%	**33**%
ru	mBERT	68%	54%	54%	60%	28%
	XLMR	**75**%	**57**%	**64**%	**73**%	31%
	mDBERT	62%	**62**%	56%	54%	31%
bg	mBERT	67%	38%	61%	52%	**38**%
	XLMR	**74**%	56%	**68**%	**64**%	27%

Table 10: NLI task. Validation accuracy. We omitted training on Slovenian since that dataset consists only of 306 items.

As we have highlighted in our analysis, the quality of cross-lingual transfer learning depends on various factors, including the pre-training and distillation details of the model, the similarity between the languages, the size of the dataset for the downstream task, etc. These observations are consistent with the conclusions of previous studies (check (Doddapaneni et al., 2021) for more details). The general observations from these experiments are:

- the multilingual pre-training is the key factor for successful cross-lingual transfer learning;
- the distillation significantly decreases the quality of cross-lingual transfer learning;
- the cross-lingual transfer quality seems to be correlated with the degree of language affinity;
- XLM-RoBERTa generally dominates over the rest of the tested models.

We are considering two major directions for future work in the scope of better understanding cross-lingual abilities of the modern models in the context of less-resourced Eastern European languages:

- leveraging the synthetic datasets and data augmentation methods for supporting the less-resourced languages;
- and evaluation of the less general models (for example, models like CroSloEngual(Ulcar and Robnik-Sikonja, 2020), BERTić (Ljubesic and Lauc, 2021), SlavicBERT (Arkhipov et al., 2019), or LitLat BERT).

[17] https://wandb.ai/eenlp/nli

6. Acknowledgements

This work was initially started as an EEML 2021 summer school project. We would like to thank all the organizers, especially Viorica Patraucean and Razvan Pascanu for their valuable support. We also would like to thank our mentor Tomáš Mikolov for wise advice and recommendations.

7. Bibliographical References

Arkhipov, M., Trofimova, M., Kuratov, Y., and Sorokin, A. (2019). Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy, August. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 09.

Avetisyan, K. and Ghukasyan, T. (2019). Word embeddings for the Armenian language: Intrinsic and extrinsic evaluation. *CoRR*, abs/1906.03134.

Batanović, V., Furlan, B., and Nikolić, B. (2011). A software system for determining the semantic similarity of short texts in Serbian. In *2011 19thTelecommunications Forum (TELFOR) Proceedings of Papers*, pages 1249–1252.

Biewald, L. (2020). Experiment tracking with Weights and Biases. Software available from wandb.com.

Butnaru, A. M. and Ionescu, R. T. (2019). MOROCO: The Moldavian and Romanian Dialectal Corpus. *CoRR*, abs/1901.06543.

Bučar, J., Žnidaršič, M., and Povh, J. (2018). Annotated news corpora and a lexicon for sentiment analysis in Slovene. *Language Resources and Evaluation*, 52, 02.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Doddapaneni, S., Ramesh, G., Khapra, M. M., Kunchukuttan, A., and Kumar, P. (2021). A primer on pretrained multilingual language models.

Dumitrescu, S. D., Rebeja, P., Lorincz, B., Gaman, M., Avram, A., Ilie, M., Pruteanu, A., Stan, A., Rosia, L., Iacobescu, C., Morogan, L., Dima, G., Marchidan, G., Rebedea, T., Chitez, M., Yogatama, D., Ruder, S., Ionescu, R. T., Pascanu, R., and Patraucean, V. (2021). LiRo: Benchmark and leaderboard for Romanian language tasks.

Emelin, D. and Sennrich, R. (2021). Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic BERT sentence embedding.

Habernal, I., Ptáček, T., and Steinberger, J. (2013). Sentiment analysis in Czech social media using supervised machine learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–74, Atlanta, Georgia, June. Association for Computational Linguistics.

Hladek, D., Stas, J., and Juhar, J. (2014). The Slovak categorized news corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1705–1708, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July. Association for Computational Linguistics.

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.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, B., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J., Wu, W., Liu, S., Yang, F., Majumder, R., and Zhou, M. (2020). XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *CoRR*, abs/2004.01401.

Ljubesic, N. and Lauc, D. (2021). Bertić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *CoRR*, abs/2104.09243.

Malajyan, A., Avetisyan, K., and Ghukasyan, T.

(2020). ARPA: Armenian paraphrase detection corpus and models. pages 35–39, 09.

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., and Augenstein, I. (2020). Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online, November. Association for Computational Linguistics.

Pasini, T., Raganato, A., and Navigli, R. (2021). XL-WSD: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proc. of AAAI*.

Pelicon, A., Pranjić, M., Miljković, D., Škrlj, B., and Pollak, S. (2020). Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17).

Piskorski, J., Laskova, L., Marcińczuk, M., Pivovarova, L., Přibáň, P., Steinberger, J., and Yangarber, R. (2019). The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 63–74, Florence, Italy, August. Association for Computational Linguistics.

Pollak, S., Purver, M., Shekhar, R., Freienthal, L., Kuulmets, H.-A., and Krustok, I. (2021). Latvian Delfi article archive (in Latvian and Russian) 1.0. Slovenian language resource repository CLARIN.SI.

Ponti, E. M., Glavas, G., Majewska, O., Liu, Q., Vulic, I., and Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning. *CoRR*, abs/2005.00333.

Purver, M., Pollak, S., Freienthal, L., Kuulmets, H.-A., Krustok, I., and Shekhar, R. (2021). Ekspress news article archive (in Estonian and Russian) 1.0. Slovenian language resource repository CLARIN.SI.

Raganato, A., Pasini, T., Camacho-Collados, J., and Pilehvar, M. T. (2020). XL-WiC: A multilingual benchmark for evaluating semantic contextualization. *CoRR*, abs/2010.06478.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., and Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Rybak, P., Mroczkowski, R., Tracz, J., and Gawlik, I. (2020). KLEJ: Comprehensive benchmark for Polish language understanding. *CoRR*, abs/2005.00630.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Scherrer, Y. (2020). TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France, May. European Language Resources Association.

Shavrina, T., Fenogenova, A., Emelyanov, A. A., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., and Evlampiev, A. (2020). RussianSuperGLUE: A russian language understanding evaluation benchmark. *CoRR*, abs/2010.15925.

Tikhonov, A. and Ryabinin, M. (2021). It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *CoRR*, abs/2106.12066.

Ulcar, M. and Robnik-Sikonja, M. (2020). FinEst BERT and CroSloEngual BERT: less is more in multilingual models. *CoRR*, abs/2006.07890.

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, Brussels, Belgium, November. Association for Computational Linguistics.

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

Yamshchikov, I. P., Shibaev, V., Khlebnikov, N., and Tikhonov, A. (2021). Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14213–14220, May.

Yang, Y., Zhang, Y., Tar, C., and Baldridge, J. (2019). PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China, November. Association for Computational Linguistics.

Žagar, A., Robnik-Šikonja, M., Goli, T., and Arhar Holdt, Š. (2020). Slovene translation of SuperGLUE. Slovenian language resource repository CLARIN.SI.