Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 470–480
https://doi.org/10.22364/bjmc.2022.10.3.19

Estonian Language Understanding: a Case Study on the COPA Task

Hele-Andra KUULMETS, Andre TÄTTAR, Mark FISHEL

University of Tartu, Ülikooli 18, 50090 Tartu, Estonia

{hele-andra.kuulmets,andre.tattar,fishel}@ut.ee

Abstract. The lack of Estonian NLU datasets severely affects advancing Estonian-specific NLP research. With this paper we aim to relieve the issue by publishing a new Estonian NLU dataset EstCOPA. We benchmark the task on several Estonian and multilingual transformer based language models, including a novel Estonian-centric GPT (GPT4Est). Moreover, we evaluate different low-cost alternatives for creating training and test datasets and outline strategies for future Estonian language understanding research.

Keywords: benchmark dataset, natural language understanding, language modeling

1 Introduction

The success of using large pre-trained language models in solving various natural language understanding (NLU) problems has motivated the creation of numerous novel benchmark datasets for NLU. However, many such datasets, such as common sense reasoning, causal reasoning and entailment recognition, to name a few, are not available for Estonian. This has become a serious bottleneck in advancing research related to Estonian language understanding. On one hand, collecting data for the tasks calls for significant amount of labour and is therefore very expensive. On the other hand, machine translation, although much cheaper, is is still not considered a worthwhile alternative.

In this paper, we introduce a new benchmark dataset for Estonian, EstCOPA which is a machine-translated and manually verified version of a well-known English causal reasoning task COPA (Roemmele et al., 2011). We use EstCOPA to evaluate mono- and multilingual transformer-based language models that are fully or partly pre-trained on Estonian. Additionally, we investigate low-cost alternatives to expensive human verified data sources. For that purpose, we fine-tune the models with different training strategies that include using synthetic training data and combining synthetic data with English data. Our results show that synthetically increasing the size of training data is

more beneficial than creating small and expensive human validated datasets. Similarly, less expensive machine translated and post edited test data is competitive with professionally translated test data.

Our contributions are summarized in the following list:

- we publish a new NLU dataset for Estonian
- we benchmark the dataset on Estonian transformer-based language model
- we train a causal Estonian-centric GPT model from scratch to complement the existing LMs for our experiments and apply it to the newly created benchmark
- we investigate low-cost methods for creating training and test data and outline strategies for future research

2 COPA Task

The Choice of Plausible Alternatives (or COPA; Roemmele et al., 2011) is a well-studied causal reasoning task for English. Given a premise and two alternatives, the task is to select the alternative that more plausibly is either the cause or effect of the premise. Table 1 presents an example.

The task is also a part of SuperGlue (Wang et al., 2019) which is a benchmark for general-purpose language understanding for English. The version of dataset included in SuperGlue has 400 train, 100 validation and 500 unlabeled test samples. The estimated accuracy of non-expert humans on the task is 100% (Wang et al., 2019) while always predicting majority class gives the accuracy of 50% as the labels in test set are split equally.

Premise: *The woman repaired her faucet.*	Choice 1: *The faucet was leaky.*
Question: *What is the cause for this?*	Choice 2: *The faucet was turned off.*
Premise: *My favorite song came on the radio.*	Choice 1: *I covered my ears.*
Question: *What happened as a result?*	Choice 2: *I sang along to it.*

Table 1: Train set examples of the English COPA task

2.1 XCOPA

There is a cross-lingual version of COPA, namely XCOPA (Ponti et al., 2020), which contains 11 languages, including Estonian. However, since XCOPA is dedicated to cross-lingual research, it only provides translations for test and validation datasets. For each target language they carefully chose a human translator who first re-annotated and then translated the datasets so that the correctness of the labelling was preserved in translation. As final labels they use majority labels obtained from 11 annotation sets.

The authors of XCOPA use the task to evaluate the ability of pre-trained multilingual language models to transfer knowledge about solving the task in English to other

languages which the models have not seen during fine-tuning. Their results show that cross-lingual transfer performs rather poorly when compared to translation-based transfer. On Estonian, the translation-based transfer achieved accurary of 81% while the best result with multilingual model transfer on Estonian was 71.4%. Additionally, Lin et al. (2021) employ XCOPA to study zero- and few-shot learning capabilities of their multilingual autoregressive language models (XGLM).

Finally, since XCOPA is also included in the cross-lingual transfer evaluation benchmark XTREME-R (Ruder et al., 2021), it attracts larger audiences of researchers who focus on cross-lingual generalization to evaluate their models on Estonian.

2.2 EstCOPA

Our work extends the work of Ponti et al. (2020) and Lin et al. (2021) - we use Estonian COPA to evaluate monolingual Estonian language models on the causal reasoning task. For that purpose, we also translate COPA training dataset to Estonian. Moreover, we re-translate validation and test sets to Estonian.[1] Our translation approach is different from Ponti et al. (2020) - we use machine translation with post-editing which is more cost efficient. This allows us to investigate the effect of different translation strategies on the results. Examples of Estonian samples are shown in Table 2.

Premise: *Naine parandas kraani.*	Choice 1: *Kraan lekkis.*
Question: *Mis on selle põhjus?*	Choice 2: *Kraan oli kinni keeratud.*
Premise: *Raadiost tuli mu lemmiklaul.*	Choice 1: *Ma katsin oma kõrvad.*
Question: *Mis juhtus selle tulemusena?*	Choice 2: *Ma laulsin kaasa.*

Table 2: Train set examples of Estonian COPA task

The datasets were translated to Estonian using the MTee machine translation system (Bergmanis et al., 2022). The translations were then corrected by a human post-editor with instructions to make the translations sound more natural, ensure the consistency of grammatical tense and consider the context of other sentences. The editor was also shown the question but not the label. Finally, we calculated translation error rate (or TER; Snover et al., 2006) to get a better understanding about the quality of the translations. Our obtained TER=0.26 indicates that the initial translations were already good when compared to post-edited data. This is not surprising since COPA consists of very short sentences.

To check that the labelling was preserved we asked the editor to re-annotate the translated validation set. Our re-annotations agreed with the original annotations in 99% of cases. We assume a similar agreement rate for test and train sets.

[1] The datasets are available at `https://huggingface.co/datasets/tartuNLP/EstCOPA`

3 Experiments

Next we perform an experimental evaluation of the newly collected data. Our goal is to cover

- masked language models (or text encoders), applied to EstCOPA task via fine-tuning large pre-trained LMs
- causal language models (or text generators), applied to EstCOPA via the zero-shot approach, without any fine-tuning
- both Estonian-centric language models as well as multilingual ones

Since a causal Estonian-centric language model does not exist yet, we created one as part of this work.

Secondly we are interested in comparing low-cost alternatives to expensive human validated training data, such as using only machine-translated data and combination of English and machine translated Estonian data. We report results on both EstCOPA and XCOPA test datasets to see the effect of translation strategy during inference.

3.1 Data setup

We use the following sources for training data: **1)** machine translated data; **2)** machine translated and post edited data, and **3)** machine translated data mixed with the original English data. While the second option is most expensive to create, it also has the best quality. The creation cost of other two datasets is much lower.

In order to measure the accuracy we use the following test sets: **1)** machine translated and post edited test data created by us, and **2)** human translated test data by Ponti et al. (2020).

3.2 Fine-tuning Models

For fine-tuning, we evaluate the following pre-trained encoder LMs:

1. XLM-RoBERTa (Conneau et al., 2020), a multilingual language model recommended as a baseline model for Estonian NLP tasks by Kittask et al. (2020). We conduct experiments with both the base model (XLM-R) and the large model (XLM-R-L).
2. EstBERT, which is a monolingual Estonian language model shown to perform better or on par with XLM-R on various tasks (Tanvir et al., 2021). The original EstBERT is trained on data with sequence length of 128 tokens. However, we also evaluate the version of the model which is trained on data with sequence length of 512 tokens (EstBERT$_{512}$).

We formulate COPA as a multiple choice classification task. Let p be premise and c_i a choice where $i \in \{1, 2\}$. For each choice c_i, the input to the encoder is then a concatenation of premise and choice in a format of p, `sest` c_i[2] if the question asks the cause

[2] In Estonian `sest` means `because`

and in a format of c_i , sest p if the question asks the effect. The classification token of the output is then projected into a score $\hat{y_i}$. The model is trained by minimizing the cross-entropy loss of a vector $\hat{\mathbf{y}}_\mathbf{i}= [\hat{y_1}, \hat{y_2}]$.

We considered learning rates $\in \{3e-6, 5e-6, 1e-5, 3e-5, 5e-5\}$ and batch sizes $\in \{8, 16\}$ as in Kittask et al. (2020), while warmup ratio and weight decay were fixed to 0.15 and 0.1 during hyperparameter search. After determining the best set of hyperparameters for each model we trained the models for 10 epochs with early stopping based on evaluation accuracy. Each model was trained from 10 random initializations. Reported results base on an ensemble of 5 best-performing models that were stopped early after the epoch with highest average evaluation accuracy.

3.3 Zero-shot Application

For zero-shot experiments we also test a multilingual and an Estonian-centric causal LM:

1. XGLM (Lin et al., 2021), a multilingual causal language model pre-trained on 30 languages, including Estonian
2. an Estonian-centric causal language model, created as part of this work: GPT4Est

For zero-shot setup we use the same input as in case of fine-tuning (p, sest c_i or c_i, sest p, depending on the question). We condition the model on the first half of the sentence and calculate normalized probabilities for both completions as described in Brown et al. (2020). For the output we pick the choice that results in a higher completion probability.

GPT4Est An Estonian-centric causal language model was missing, thus we trained it as part of our work. We used a total of 2.2 billion token monolingual corpus for the task, which included the Estonian National Corpus 2019 (Koppel and Kallas, 2020), the Estonian portion of Common Crawl[3] and Estonian News Crawl (Akhbardeh et al., 2021). Differently from the general approach to GPT-style model creation we prepended each document with a tag, specifying the source of the data: wiki, news, general, articles or web.

The implementation was taken from HuggingFace[4]. We trained two versions of the model, base[5] and large[6]. Their differences are summarized in Table 3. Both models were trained starting from random initialization with default GPT-2 training settings. Training was continued for 3 epochs for both model sizes. For the large model this took 19 days on 4 A100 GPUs with 80 GB of VRAM with the maximum possible batch size; the base model took 9 days.

[3] https://commoncrawl.org
[4] https://huggingface.co/gpt2
[5] https://huggingface.co/tartuNLP/gpt-4-est-base
[6] https://huggingface.co/tartuNLP/gpt-4-est-large

Parameter	Base	Large
#Layers	12	24
#Heads	12	16
Embedding size	768	1536
Context length	1024	1024
Total #params	118.7M	723.6M

Table 3: Differences between GPT4Est base and large

4 Results

4.1 Fine-tuned models

We first present the results obtained by fine-tuning pretrained encoders (Table 4). Overall, the accuracies on test dataset are, with one exception, rather poor. For the comparison, we have included cross-lingual transfer results reported by Ponti et al. (2020) to the table. The results indicate that language-specific approaches might not give any benefits over transfer-based methods in case of very small datasets. Moreover, there's also no constant improvement in results when comparing machine translated training data with post edited training data.

However, we do observe constant improvement when mixing machine translated data with original English training data, even for monolingual language models. Our best model improves over our second best model with 15.2 percentage points. We conclude that in scenarios with limited data available, the size of the training dataset is more important than the quality and synthetically increasing the size of the training dataset is more beneficial in terms of both, performance and cost, than careful human validation of train samples.

	our models			Ponti et al. (2020)	
	MT	PE	MT+En	En	En_{TLV}
EstBERT	53.2	57.2	**58.6**	-	-
EstBERT_{512}	55.8	55.0	**57.0**	-	-
XLM-R	53.4	53.6	56.6	**59.8**	57.8
XLM-R-L	57.4	55.4	**73.8**	49.6	49.4

Table 4: Accuracy on XCOPA test dataset; **MT**: machine translated training data; **PE**: machine translated and post edited training data; **MT+En**: machine translated data mixed with English training data; **En**: English training data; $\textbf{En}_{\textbf{TLV}}$: English training data and machine translated English validation data. Best accuracy reported by Ponti et al. (2020) for Estonian was 71.4%, obtained with using a large auxiliary dataset (33K instances) for training and machine translated English validation data.

4.2 Zero-shot results

We present our zero-shot results in Table 5. We prepended our input to GPT4Est with a domain tag as was done during pretraining of that model. We tested every tag as well as using no tag and report the best obtained accuracy. The input to XGLM was not modified in any way. In our experiments, the best performing model is a 7.5 billion parameter XGLM model.

XGLM			GPT4Est		
size	XCOPA	EstCOPA	size	XCOPA	EstCOPA
546M	52.8	52.4	base	53.0	49.4
1.7B	50.8	50.2	large	53.0	51.4
2.9B	51.4	49.8			
4.5B	51.0	51.6			
7.5B	**54.2**	**53.4**			

Table 5: Zero-shot accuracies on XCOPA and EstCOPA test datasets. For the comparison, Lin et al. (2021) reported zero-shot accuracy of 61.6% with XGLM 7.5B parameter model on Estonian, however, they used a different prompt.

4.3 Comparison of test datasets

We use our best-performing fine-tuned models to report the results on our own test dataset (Table 6). We observe constant increase in test accuracy with average increase of 1.12 percentage points when compared to XCOPA test results. This suggests that the task is slightly easier to solve for the model on EstCOPA test set, however, there is still a lot of room for the improvement.

	EstBERT	EstBERT_{512}	XLM-R	XLM-R-L
XCOPA	58.6	57.0	56.6	73.8
EstCOPA	59.0	59.0	57.2	75.6

Table 6: Comparison of accuracies of **MT+En** models on XCOPA and EstCOPA test datasets.

Table 5 shows that for causal language models, on the other hand, EstCOPA is harder to solve as the accuracy is almost always slightly worse. For XGLM the average difference is 0.8 percentage points, and for GPT4Est it's 2.6%. In case of GPT4Est, we used the same prefix for EstCOPA as for XCOPA. Without any domain tags on both datasets the difference would have been 0.3% indicating that the tag has a strong and unpredictable influence on the generated output.

5 Discussion

Our work focuses on efficiently solving Estonian NLU tasks by making use of English-to-Estonian machine translation. Alternatively, one could take advantage of Estonian-to-English translation systems. Although this direction is not covered in this work, some results on XCOPA have been reported by other authors and are summarized in Table 7. As the table shows, simple English baselines are surprisingly good. Moreover, zero- and few-shot results on translated test data are close to our best-performing model.

Setup	**Model**	**Accuracy**
Ponti et al. (2020)		
Train on en, translate test data to en	XLM-R-L	76.8
Train on en, translate test data to en	RoBERTa-L	81.0
Lin et al. (2021)		
Zero-shot, translate test data to en	$\text{XGLM}_{6.7\text{B}}$ *en*-only	72.4
4-shot, translate test data to en	$\text{XGLM}_{6.7\text{B}}$ *en*-only	73.6

Table 7: Accuracies of fine-tuned and zero-shot models on translated XCOPA test set. Models by Ponti et al. (2020) are fine-tuned using large auxiliary dataset (33K instances) and validated on translated English validation set.

Isbister et al. (2021) raise a provocative question whether native non-English language models should be trained at all if machine translation can be used instead. We believe that answering to this question, at least for Estonian, requires further work on the topic and more careful experimentation with translation directions, strategies and datasets. Moreover, assuming the high quality of machine translated data, the comparison between English and native models is often unfair, as English models have been fine-tuned on larger datasets.

As for EstCOPA, no additional training data sources were used during fine-tuning but we assume that the results can be improved if larger training data is used for fine-tuning, as was by Ponti et al. (2020). However, we leave experiments with that for future work. Secondly, COPA was created by American English speakers. Therefore the sentences sometimes contain cultural context that is unfamiliar to Estonian speakers and Estonian language models. It might be that due to this context some samples are naturally easier to understand for English language models. Translating original Estonian datasets to English could lead to different results.

6 Related Work

Besides the causal reasoning task, the effect of using machine translated data on solving Estonian NLU tasks has been previously studied for extractive question-answering (Käver, 2021) and abstractive text summarization (Härm, 2021). Käver (2021) reported that translating the English training set to Estonian outperforms translating test data

to English when fine-tuning XLM-RoBERTa. However, the best results were achieved when fine-tuning the model sequentially on English data, machine translated Estonian data and task-specific Estonian data.

In addition, Härm (2021) generated best abstractive summaries with a pre-trained and fine-tuned monolingual English model that was applied to machine translated English articles. This outperformed the model initialized from EstBERT that was additionally fine-tuned on a machine translated Estonian dataset. However, the sizes of fine-tuning datasets differed significantly, making the direct comparison of the two models unfair.

Finally, Isbister et al. (2021) show that machine translation coupled with large English models outperforms native models in most Scandinavian languages and raise a question whether native language models should be trained at all.

7 Conclusion

We introduce EstCOPA, a new NLU dataset for Estonian which consists of novel Estonian train, validation and test datasets. Our validation and test sets are machine translated and post-edited cost-efficient alternatives to Estonian datasets included to XCOPA (Ponti et al., 2020), with EstCOPA's test tasks being slightly simpler to solve for the models than XCOPA's test tasks. We report baseline accuracy of EstCOPA on EstBERT and zero-shot baseline accuracy on novel GPT4Est. Additionally, we evaluate various training strategies and find that using human verified training data has no advantage over machine translated data in a low resource setting. However, combining English and machine translated Estonian constantly improves the results of our fine-tuned models. Based on these results we encourage researches working on Estonian NLU tasks not to be afraid to rely on machine translated train and test datasets as it appears to be a fast and efficient way to nurture NLU research for Estonian.

Acknowledgements

This article has been financed/supported by European Social Fund via "ICT programme" measure.

References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21), *Proceedings of the Sixth Conference on Machine Translation*, Association for Computational Linguistics, Online, pp. 1–88.
`https://aclanthology.org/2021.wmt-1.1`

Bergmanis, T., Pinnis, M., Rozis, R., Šlapiņš, J., Šics, V., Bernāne, B., Pužulis, G., Titomers, E., Tättar, A., Purason, T., Kuulmets, H.-A., Luhtaru, A., Rätsep, L., Tars, M., Laumets-Tättar, A., Fishel, M. (2022). MTee: Open machine translation platform for Estonian government, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, European Association for Machine Translation, Ghent, Belgium, pp. 309–310. https://aclanthology.org/2022.eamt-1.44

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. (2020). Language models are few-shot learners. https://arxiv.org/abs/2005.14165

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 8440–8451. https://aclanthology.org/2020.acl-main.747

Härm, H. (2021). *Abstractive summarization of news broadcasts for low resource languages*, Master's thesis, Tallinn University of Technology.

Isbister, T., Carlsson, F., Sahlgren, M. (2021). Should we stop training more monolingual models, and simply use machine translation instead?, *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), pp. 385–390. https://aclanthology.org/2021.nodalida-main.42

Kittask, C., Milintsevich, K., Sirts, K. (2020). Evaluating multilingual bert for estonian, *Volume 328: Human Language Technologies – The Baltic Perspective* pp. 19–26.

Koppel, K., Kallas, J. (2020). Eesti keele ühendkorpus 2019.

Käver, A. (2021). *Extractive question answering for estonian language*, Master's thesis, Tallinn University of Technology.

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., Stoyanov, V., Li, X. (2021). Few-shot learning with multilingual language models. https://arxiv.org/abs/2112.10668

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 2362–2376. https://aclanthology.org/2020.emnlp-main.185

Roemmele, M., Bejan, C. A., Gordon, A. S. (2011). Choice of plausible alternatives: An evaluation of commonsense causal reasoning, *2011 AAAI Spring Symposium Series*.

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 10215–10245. https://aclanthology.org/2021.emnlp-main.802

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006). A study of translation edit rate with targeted human annotation, *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Association for Machine Translation

in the Americas, Cambridge, Massachusetts, USA, pp. 223–231.
https://aclanthology.org/2006.amta-papers.25

Tanvir, H., Kittask, C., Eiche, S., Sirts, K. (2021). EstBERT: A pretrained language-specific BERT for Estonian, *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), pp. 11–19.
https://aclanthology.org/2021.nodalida-main.2

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. R. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems, *arXiv preprint 1905.00537* .

Received August 19, 2022 , accepted August 27, 2022