# SLOVAKT5-SMALL

**Richard Cepka**
Comenius University in Bratislava

**ABSTRACT**

Low resource languages are missing the pre-trained language models, and they are under-researched from an NLP point of view. That's why we trained and explored SlovakT5-small and filled the gap of missing pre-trained encoder-decoder architecture in the Slovak language.

***K*eywords** low resource languages, transfer learning, natural language processing, multi-task learning, attention based models, deep learning

## 1 Introduction

Recently was published pre-trained BERT model in the Slovak language, called SlovakBERT [1], and his distilled version [2]. Also in HugginFace is availible pre-trained Slovak GPT-J [3]. The last missing piece to the complete family of standard transformer architectures (encoder, decoder and encoder-decoder) is T5 [4] style, encoder-decoder model.

T5 was proposed in 2020 as a general language modelling framework to view each task as a language modelling problem. It was pre-trained on a multi-task mixture of unsupervised and supervised tasks and works well on a variety of tasks out-of-the-box by prepending a different prefix to the input corresponding to each task, e.g., for translation: translate English to German: ..., for summarization: summarize: ....

Due to a lack of computation resources (only Google Colab), we can't afford to train T5 from scratch (including the 60M version). That is why we use multilingual[1] mT5-small [5] as starting point. We believe language understanding of the mT5 model can speed up the training of the pure Slovak model.[2]

## 2 Data

The big challenge of training a low resource language model is the lack of data, specially supervised data for the downstream tasks. One possible solution is to automatically translate the English benchmark dataset by some translation model. This approach is not perfect, and translated dataset may contain some artefacts. In this section, we summarise the Slovak language dataset used in our project.

### 2.1 Pre-training Data set

**mC4-sk (400K)** A multilingual colossal, cleaned version of Common Crawl's web crawl corpus. Include 108 languages. **We use first 400 000 samples from Slovak language part.**

**OSCAR-sk** OSCAR or Open Super-large Crawled ALMAnaCH coRpus is a huge multilingual corpus obtained by language classification and filtering of the Common Crawl corpus using the goclassy architecture. Include 166 languages. **We use whole Slovak language part.**

### 2.2 Evaluation Data set

**SST2-sk** Sentiment analysis dataset, **binary classification task: positive sentiment, negative sentiment**. We obtained this dataset from SlovakBERT Auxiliary Repository [6]. It includes reviews from 7 categories with positive,

[1] The Slovak language was also included in the training corpus.

[2] The pre-trained model can find at `https://huggingface.co/ApoTro/slovak-t5-small`.

neutral and negative sentiment labels. After filtering out reviews with neutral labels (only 57 samples), we obtained 620 reviews with balanced classes. We further split these 620 reviews into **train: 372, validation: 99 and test: 149** subsets.

**STSB-sk** Sentence similarity dataset contains **two sentences with a floating-point number between** 0 **and** 5 **as a target**, where the highest number means higher similarity [7]. The Slovak version of this dataset [8] was obtained by Ivan Agarský due translation, using the English-Slovak translation model *opus-mt-en-sk* by Helsinki-NLP [9]. The dataset contains **train: 5 749, validation: 1 500 and test: 1 379** examples.

**BoolQ-sk** In the Boolean questions dataset each entry contains a **text passage, a question related to that passage and a yes/no answer to this question** [10]. The Slovak version of this dataset [11] was obtained by Ivan Agarský due translation, using the English-Slovak translation model *opus-mt-en-sk*. We split this dataset to **train: 4 735, validation: 510 and test: 1 190** subsets.

## 3 Approach

In this section, we describe our approach to training SlovakT5-small.

### 3.1 Tokenizer

We trained Sentence Piece Unigram Tokenizer on mC4-sk (400K) and OSCAR-sk with a vocabulary size 32 000. After that we add 100 sentinel tokens, as it was propose in T5 paper for unsupervised training.

### 3.2 Embedding pre-training

Pre-trained mT5-small has a huge vocabulary size of 250 112. To save memory we randomly reinitialize the embedding matrix and langue modelling head to size (512, 32 100), and further to save memory, we tie parameters of the embedding matrix and language modelling head.

Then we freeze all parameters except for the embedding matrix (tie with language modelling head) and fine-tune it on mC4-sk (400K) with T5 unsupervised objective for 1 epoch.

The motivation behind this approach is that mT5-small was trained on mC4, and due to training only embedding matrix on mC4-sk (400K) we tried to match Slovak language knowledge of mT5-small without catastrophic forgetting.

### 3.3 Final training

After previous two steps we trained end-to-end SlovakT5-small on whole OSCAR-sk dataset with T5 unsupervised objective. After 3 epochs[3] on OSCAR-sk dataset we achieved on evaluation set (randomply sampled 5% of OSCAR-sk dataset) loss and accuracy of 2.43 and 55.8 respectively.[4]

### 3.4 Evaluation

We evaluated SlovakT5-small on 3 tasks: SST2-SK, STSB-SK and BoolQ-SK, where each task we reformulated as a language modelling task, the same as in the T5 paper. We translated prefixes to the Slovak, see Appendix.

For each task, we trained with a learning rate of 1e-4 and batch size 24. We set a number of epochs by the best score on the validation set. After that, we trained the model on the merged train and validation set and reported metrics on the test set.

## 4 Results

If we take into account model size and number of Slovak data seen during pre-training, SlovakT5-small end up comparable to other models. We can also see that SlovakT5-small slightly outperform DistilSlovakBERT on the BoolQ-sk. This can be due to the nature of the BoolQ-SK dataset (translated from English to Slovak).

[3] In the last third epoch, the improvement was very tiny. Improvement was only 1.8 accuracy on the evaluation set.

[4] For comparison, trained T5-base after 3 epochs achieved 2.36 loss and 57.0 accuracy [12], on comparable dataset size (OSCAR-no).

| Model | SST2-sk | STSB-sk | | BoolQ-sk | # Params |
|---|---|---|---|---|---|
| | (Accuracy) | (Pearson) | (Spearman) | (Accuracy) | |
| SlovakBERT | 0.966 | - | 0.781 | 0.709 | 124M |
| DistilSlovakBERT | 0.859 | - | 0.734 | 0.662 | 82M |
| mT5-small | 0.570 | 0.683 | 0.675 | 0.617 | 300M |
| SlovakT5-small | 0.738 | 0.685 | 0.667 | 0.675 | 60M |

SlovakBERT and DistilSlovakBERT metrics on STSB-sk and BoolQ-sk are from [2].

## 5 Conclusion

We proposed first step to transfer T5 style models to Slovak language and also released the smallest pre-trained model from this family, SlovakT5-small.

We didn't do any ablation studies due to computation constrains. That's why in future work we encourage comparing our approach with pure training from scratch. Also, scale things ups and more extensively evaluate a trained model. Last and from our perspective most important thing is to set a proper evaluation benchmark for the Slovak language. It can hugely accelerate research of these language models in low-resource languages.

## References


[1] Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. Slovakbert: Slovak masked language model. *arXiv:2109.15254*, 2021.

[2] Ivan Agarský. Distilling the knowledge of slovakbert. 2022.

[3] Milos Kondela. Slovak GPT-J-1.4B. `https://huggingface.co/Milos/slovak-gpt-j-1.4B`, 2022.

[4] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[5] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2020.

[6] Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marián Šimko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. Slovakbert auxiliary repository. `https://github.com/kinit-sk/slovakbert-auxiliary/tree/main/sentiment_reviews`, 2021.

[7] Philip May. Machine translated multilingual sts benchmark dataset. `https://huggingface.co/datasets/stsb_multi_mt`, 2021.

[8] Ivan Agarský. Stsb-sk. `https://huggingface.co/datasets/crabz/stsb-sk`, 2022.

[9] Helsinki-NLP. opus-mt-en-sk. `https://huggingface.co/Helsinki-NLP/opus-mt-en-sk`, 2020.

[10] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019.

[11] Ivan Agarský. Boolq-sk. `https://huggingface.co/datasets/crabz/boolq_sk`, 2022.

[12] Patrick von Platen. Language model training examples. `https://github.com/huggingface/transformers/tree/main/examples/flax/language-modeling`.

# Appendix

## Reformulation to language modeling

### SST2-sk

**Original input:** `Trochu mi tu chýba konzistentnosť kvality fotiek.`

**Processed input:** `sst2 veta: Trochu mi tu chýba konzistentnosť kvality fotiek.`

**Original target:** `0`

**Processed target:** `negatívna`

### STSB-sk

**Original input:**

**Sentence 1:** `Traja muži hrajú šach.`
**Sentence 2:** `Dvaja muži hrajú šach.`

**Processed input:** `stsb veta1: Traja muži hrajú šach. veta2: Dvaja muži hrajú šach.`

**Original target:** `4.25`

**Processed target:** `4.2`

### BoolQ-sk,

**Original input:**

**Question:** `je pán prsteňov po hobite.`
**Sentence:** `Pán prsteňov je impozantný fantasy román, ktorý napísal anglický autor a učenec J. R. Tolkien. Príbeh sa začal ako pokračovanie fantasy románu Tolkiena z roku 1937 Hobbit, ale nakoniec sa vyvinul do oveľa väčšieho diela. Napísaný v etapách medzi rokmi 1937 a 1949, Pán prsteňov je jedným z najpredávanejších románov, aké kedy boli napísané, s viac ako 150 miliónmi predaných kópií.`

**Processed input:** `qnli otázka: je pán prsteňov po hobite. veta: Pán prsteňov je impozantný fantasy román, ktorý napísal anglický autor a učenec J. R. Tolkien. Príbeh sa začal ako pokračovanie fantasy románu Tolkiena z roku 1937 Hobbit, ale nakoniec sa vyvinul do oveľa väčšieho diela. Napísaný v etapách medzi rokmi 1937 a 1949, Pán prsteňov je jedným z najpredávanejších románov, aké kedy boli napísané, s viac ako 150 miliónmi predaných kópií.`

**Original target:** `1`

**Processed target:** `vyplýva`

## Training tips on Google Colab

1. Use Jax TPU support, is it way faster than GPU.
2. Preprocess and tokenize dataset before TPU training.
3. Frequently save checkpoints, e.g. on Google disk.
4. Most of the time TPU is available in the early morning or late evening.