Distilling the Knowledge of SlovakBERT

Školiteľ: prof. Ing. Igor Farkaš, Dr.
Konzultant: Mgr. Marek Šuppa
Študent: Bc. Ivan Agarský

Content

- Introduction
- Language model
- BERT and the Transformer
- SlovakBERT
- Knowledge distillation
- Experiments
- Results
- Conclusion

Introduction

- Small datasets for NLP tasks
- Large pre-trained models
- Problems on edge devices
- Knowledge distillation

Language model

- $P(\boldsymbol{w}) = P(w_1, w_2, \ldots, w_k)$
- $P(w_k \mid w_{k-1}, \ldots, w_1)$

BERT

“The incident happened [MASK] the building.”

outside	0.393
inside	0.278
in	0.207
within	0.051
…	

Devlin et. al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019

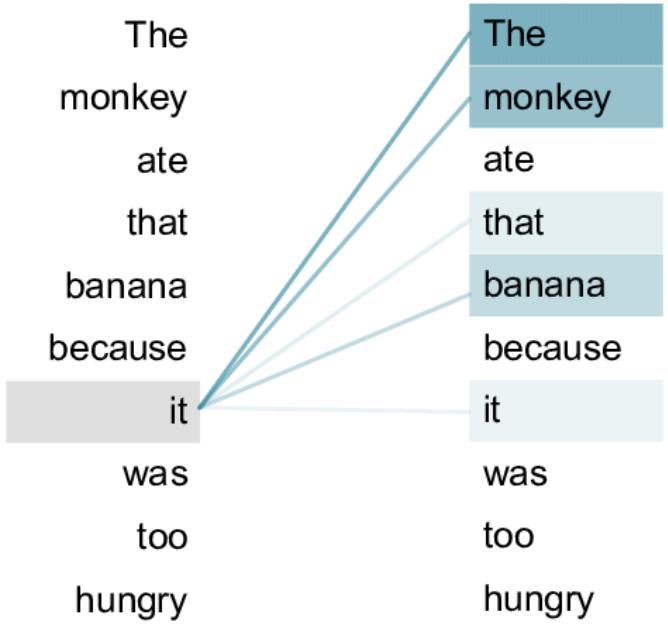
BERT

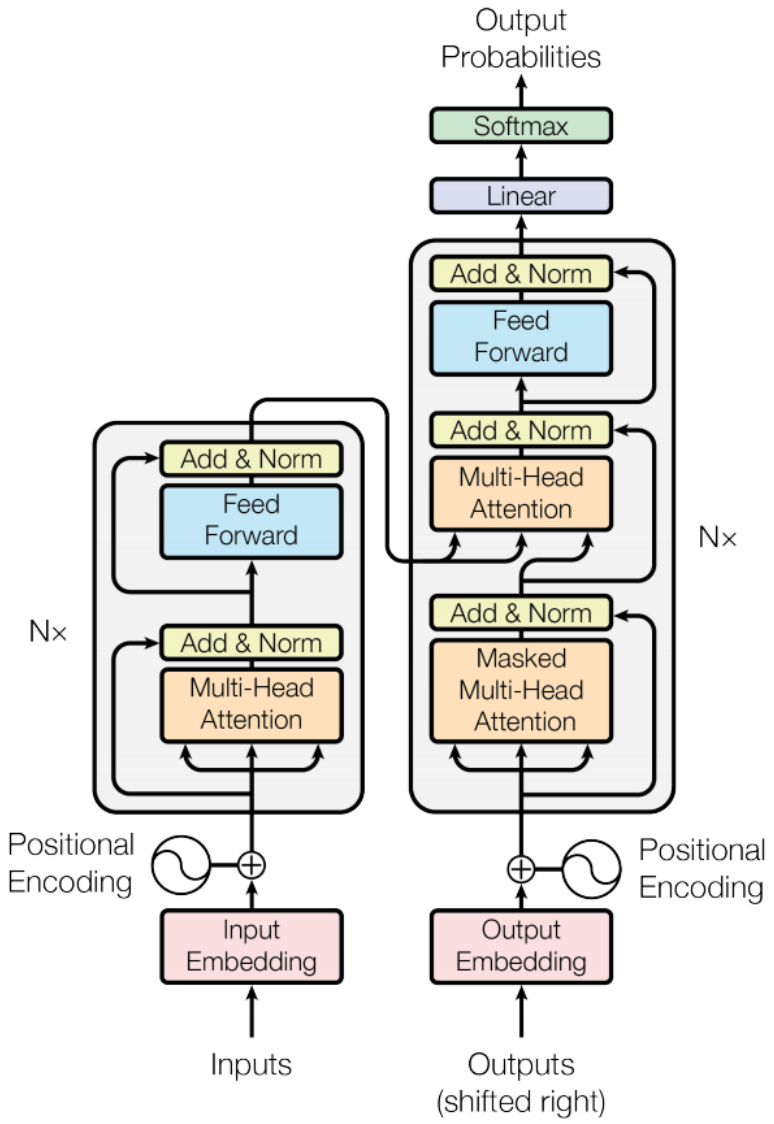
- Unsupervised pre-training:
  - Fill-mask

- Supervised fine-tuning:
  - Question answering
  - Text generation
  - Translation
  - …

The Transformer

- Neural network
- Multi-Head Attention





Vaswani et. al. *Attention is all your need*, 2017

SlovakBERT

- Slovak language model (September 2021)
- 19 GB dataset: Wikipedia, Open Subtitles, OSCAR + web crawl

Pikuliak et. al. *SlovakBERT: Slovak Masked Language Model*, 2021

Downstream tasks

- SlovakBERT:
  - Universal part-of-speech (UPOS) tagging
  - Sentiment analysis
  - Document classification
  - Semantic text similarity (STS)

Knowledge distillation

- SlovakBERT requires 476 MB (PyTorch), 627 MB (TensorFlow)

- Teacher-Student
- Logits: $\mathcal{L}_{KL}\left(P(\boldsymbol{z}_t), P(\boldsymbol{z}_s)\right)$
- Feature maps

- DistilBERT

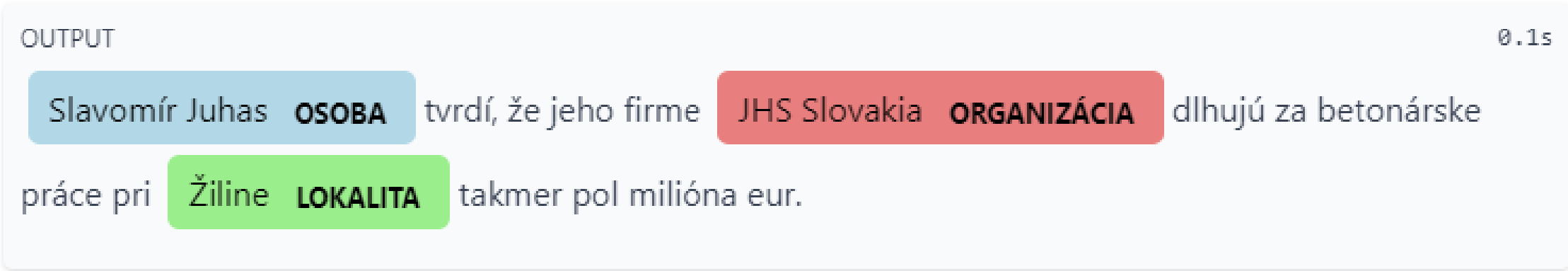
Hinton et al. *Distilling the knowledge in a neural network, 2015*

Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*, 2020

Experiments

- 6 experiments
  - 2 x 6-layer model
  - 4 x 4-layer model
- C4 dataset (subset 1.9 GB)
- Evaluation on NER, UPOS, **STS** and **BoolQ**

NER


OUTPUT
0.1s
Slavomír Juhas OSOBA tvrdí, že jeho firme JHS Slovakia ORGANIZÁCIA dlhujú za betonárske práce pri Žiline LOKALITA takmer pol milióna eur.

UPOS


Ruská ADJ armáda NOUN pokračuje VERB v ADP útokoch NOUN na ADP východe NOUN Ukrajiny PROPN . PUNCT

STS

Score	Explanation
5	The two sentences are completely equivalent, as they mean the same thing.
4	The two sentences are mostly equivalent, but some unimportant details differ.
3	The two sentences are roughly equivalent, but some important information differs/missing.
2	The two sentences are not equivalent but share some details.
1	The two sentences are not equivalent but are on the same topic.
0	The two sentences are completely dissimilar.

BoolQ

- **Passage**: *In the 1998 war film Saving Private Ryan, General George Marshall (played by Harve Presnell) reads the Bixby letter to his officers before giving the order to find and send home Private James Francis Ryan after Ryan's three brothers died in battle.*

  Question: *Did Abraham Lincoln write the letter in Saving Private Ryan?*

- Yes/No

Experiments

	KL divergence	Cross-entropy	Cosine embedding	Weight init
Experiment 1	0.625	0.25	0.125	[1, 3, 5, 8, 10, 12]
Experiment 2	0.625	0.25	0.125	[1, 2, 4, 6, 9, 11]
Experiment 3	0.6	0.2	0.2	[1, 5, 8, 11]
Experiment 4	0.7	0.2	0.1	[1, 2, 3, 4]
Experiment 5	0.7	0.2	0.1	[1, 3, 5, 7]
Experiment 6	0.7	0.2	0.1	[1, 2, 3, 4]

Experiments

(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

(d) Experiment 4

(e) Experiment 5

(f) Experiment 6

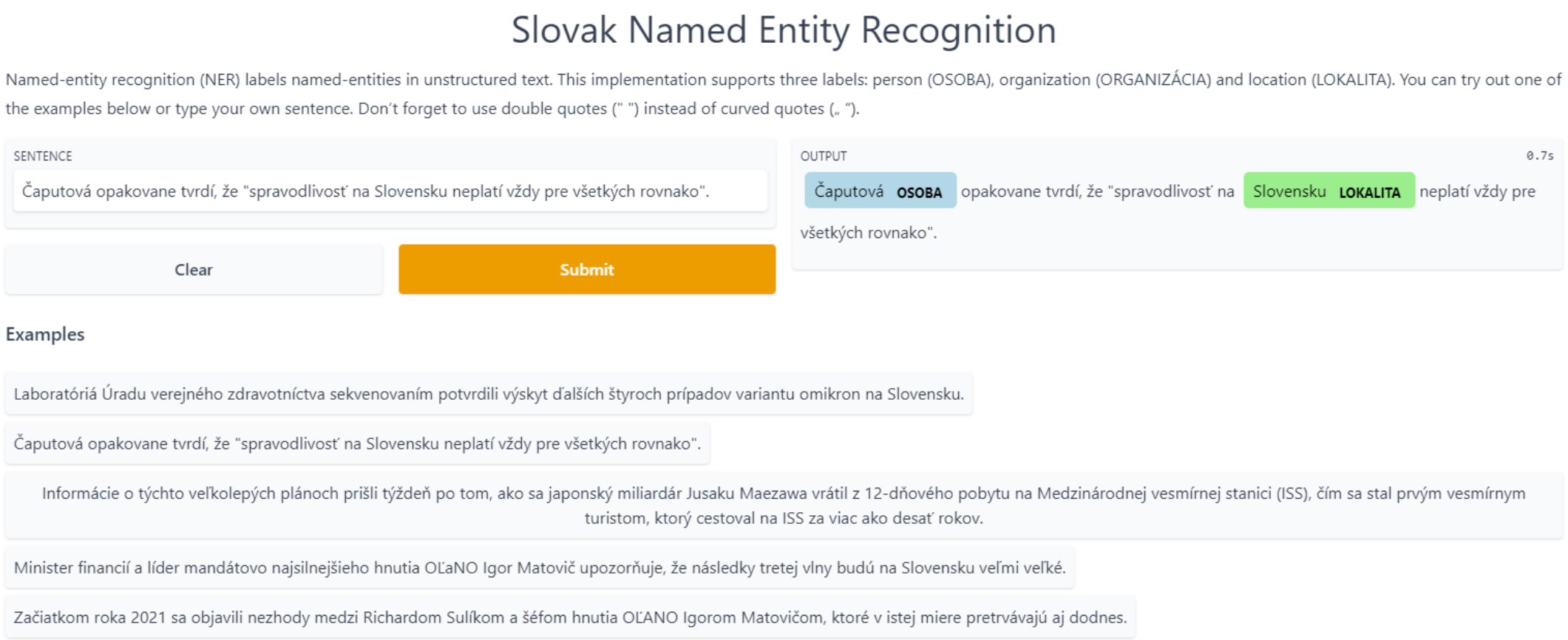
Results

Model	**NER** (Macro-F1)	**POS** (Macro-F1)	**STS** (Spearman)	**BoolQ** (Accuracy)	**# Params**
SlovakBERT	0.939	**0.983**	0.781	**0.709**	124M
FERNET-cc_sk	**0.941**	0.980	**0.788**	0.663	162M
Experiment 1	0.929	0.976	0.713	0.649	82M
Experiment 2	**0.931**	**0.979**	0.734	0.662	82M
Experiment 3	0.907	0.972	0.720	0.646	67M
Experiment 4	0.916	0.974	**0.743**	**0.668**	67M
Experiment 5	0.915	0.973	0.740	0.645	67M
Experiment 6	0.916	0.975	0.693	0.642	67M

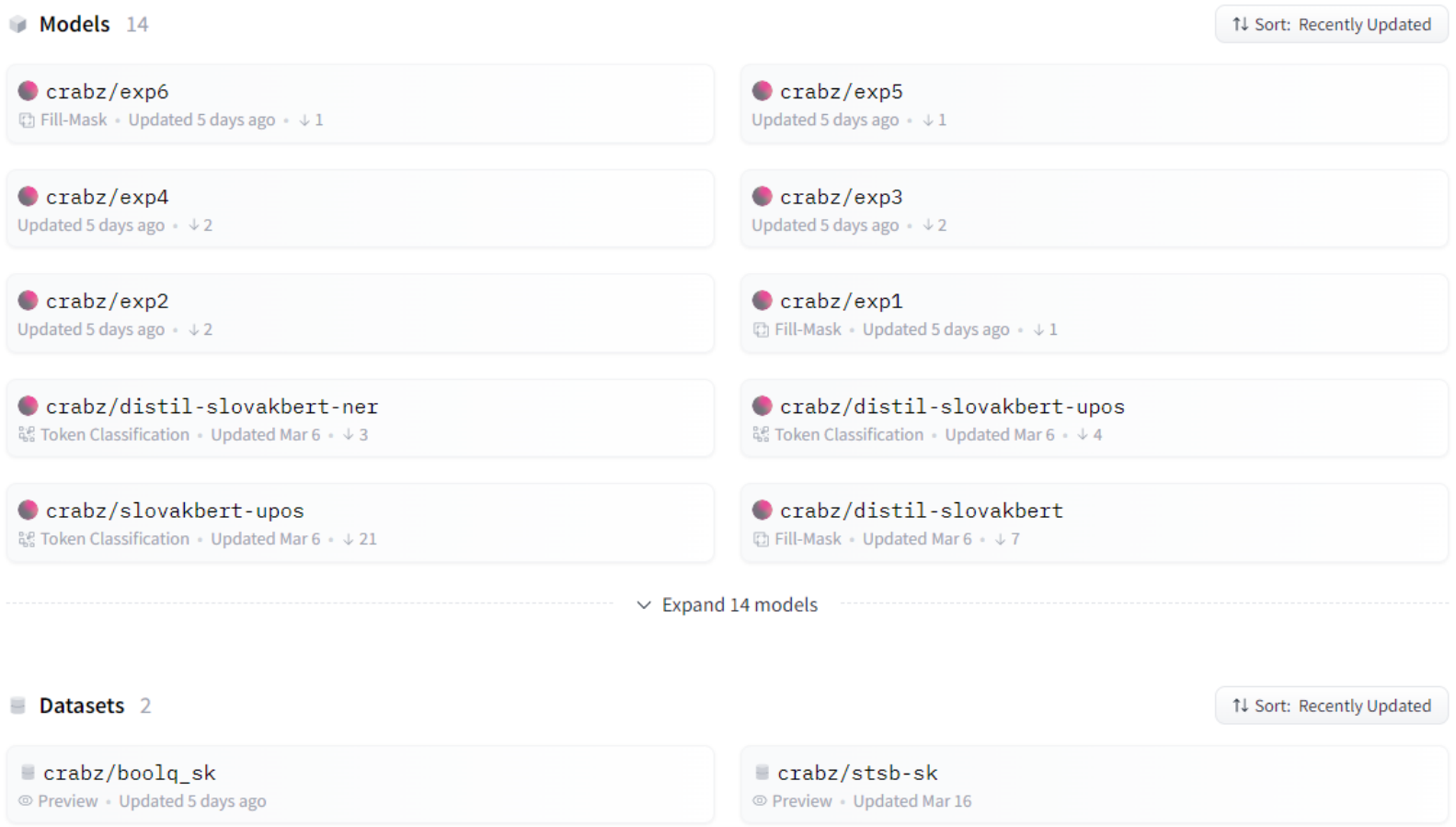
Results

- Memory requirements 476 MB -> 260 MB
- Inference time 13 ms -> 6 ms

Deployed NER model



Published models and datasets



Summary of contributions

- We achieve from 91% to 99% of the original performance on four tasks with 46% fewer parameters
- We demonstrate importance of choosing the right initial student weights
- We publish Slovak datasets for STS and BoolQ (machine-translated)
- We deploy Slovak NER and UPOS model for online inference