IS-2021
SiKDD 2021
Ljubljana, 4th October

SloBERTa
Slovene monolingual large pretrained model

Matej Ulčar and **Marko Robnik-Šikonja**
University of Ljubljana,
Faculty of Computer and Information Science

Introduction

Transformers in NLP (BERT, GPT-3, T5, etc.)

- Pre-trained to predict masked or next word, generate text, inference
- Fine-tuned on a specific task

Monolingual (initially English and Chinese)

Massive multilingual (100+ languages)

No monolingual model for Slovene language

SloBERTa

First Slovene monolingual transformer masked language model

Based on RoBERTa (only masked language model)

Architecture shared with BERT-base (RoBERTa-base) model

- 12-layer transformer encoder
- hidden layer size 768
- max sequence length 512 tokens

Datasets

Corpus	Genre	Tokens	Words
Gigafida 2.0	general language	1.33	1.11
Janes*	social media	0.10	0.08
KAS	academic	1.70	1.33
siParl 2.0	parliamentary	0.24	0.20
slWaC 2.1	web crawl	0.90	0.75
Total		4.27	3.47
Total after deduplication		4.20	3.41

Tokenizer

Slovene sentencepiece model

tokenizer, detokenizer, BPE encoder

trained on a (random) subset of the dataset

32,000 subword tokens in vocabulary

Example sentence

“Letenje je bilo predmet precej starodavnih zgodb.”

SloBERTa: '▁Le', 'tenje', '▁je', '▁bilo', '▁predmet', '▁precej', '▁staroda', 'vnih', '▁zgodb', '.'

mBERT: 'Let', '##en', '##je', 'je', 'bilo', 'pred', '##met', 'pre', '##cej', 'star', '##oda', '##vnih', 'z', '##go', '##d', '##b', '.'

Training

15% tokens in a sequence randomly masked

→ predict masked tokens!

Whole-word masking

Trained for 200,000 steps (~98 epochs)

~4 weeks on Nvidia DGX A100 using 4 GPUs

Evaluation

Five downstream tasks:

named entity recognition (NER)

part-of-speech tagging (POS)

dependency parsing (DP)

sentiment analysis (SA)

word analogies (WA) reinterpreted as masked token prediction

Compared with multilingual models:

multilingual BERT (mBERT)

XLM-RoBERTa (XLM-R)

CroSloEngual BERT (CSE-BERT) (trilingual)

Evaluation results

Model	NER	POS	DP	SA	WA
fastText	0.478	0.527	/	0.435	/
ELMo	0.849	0.966	**0.914**	0.510	/
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
CSE-BERT	0.928	0.990	0.854	0.610	0.195
SloBERTa	**0.933**	**0.991**	0.844	**0.623**	**0.405**

Evaluation

SloBERTa best on 4/5 tasks

Big improvements on WA and SA

SA limited by low annotator agreement

Model	NER	POS	DP	SA	WA
fastText	0.478	0.527	/	0.435	/
ELMo	0.849	0.966	**0.914**	0.510	/
mBERT	0.885	0.984	0.681	0.576	0.061
XLM-R	0.912	0.988	0.793	0.604	0.146
CSE-BERT	0.928	0.990	0.854	0.610	0.195
SloBERTa	**0.933**	**0.991**	0.844	**0.623**	**0.405**

ELMo surprisingly best on dependency parsing

fastText and ELMo not fully comparable with others

Conclusions

Publicly released SloBERTa model

CLARIN.SI: hdl.handle.net/11356/1397

HuggingFace: huggingface.co/EMBEDDIA/sloberta

Impoved performance on most evaluation tasks

Future work:

further evaluation on more tasks

comparative evaluation between mono- and multilingual models on more languages

train a larger Slovene transformer model