SloBERTa: Slovene monolingual large pretrained masked language model

Matej Ulčar and Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Ljubljana, Slovenia
{matej.ulcar,marko.robnik}@fri.uni-lj.si

ABSTRACT

Large pretrained language models, based on the transformer architecture, show excellent results in solving many natural language processing tasks. The research is mostly focused on English language; however, many monolingual models for other languages have recently been trained. We trained first such monolingual model for Slovene, based on the RoBERTa model. We evaluated the newly trained SloBERTa model on several classification tasks. The results show an improvement over existing multilingual and monolingual models and present current state-of-the-art for Slovene.

KEYWORDS

natural language processing, BERT, RoBERTa, transformers, language model

1 INTRODUCTION

Solving natural language processing (NLP) tasks with neural networks requires presentation of text in a numerical vector format, called word embeddings. Embeddings assign each word its own vector in a vector space so that similar words have similar vectors, and certain relationships between word meanings are expressed in the vector space as distances and directions. Typical static word embedding models are word2vec [19], GloVe [24], and fastText [1]. ELMo [25] embeddings are an example of dynamic, contextual word embeddings. Unlike static word embeddings, where a word gets a fixed vector, contextual embeddings ascribe a different word vector for each occurrence of a word, based on its context.

State of-the-art text representations are currently based on the transformer architecture [35]. GPT-2 [27] and BERT [5] models are among the first and most influential transformer models. Due to their ability to be successfully adapted to a wide range of tasks, such models are, somewhat impetuously, called foundation models [2, 17]. While GPT-2 uses the transformer's decoder stack to model the next word based on previous words, BERT uses the encoder stack to encode word representations of a masked word, based on the surrounding context before and after the word. Previous embedding models (e.g., ELMo and fastText) were used to extract word representations which were then used to train a model on a specific task. In contrast to that, transformer models are typically fine-tuned for each individual downstream task, without extracting word vectors.

Information Society 2021, 4–8 October 2021, Ljubljana, Slovenia

Successful transformer models typically contain more than 100 million parameters. To train, they require considerable computational resources and large training corpora. Luckily, many of these models are publicly released. Their fine-tuning is much less computationally demanding and is accessible to users with modest computational resources. In this work, we present the training of a Slovene transformer-based masked language model, named SloBERTa, based on a variant of BERT architecture. SloBERTa is the first such publicly released model, trained exclusively on the Slovene language corpora.

2 RELATED WORK

Following the success of the BERT model [5], many transformer-based language models have been released, e.g., RoBERTa [14], GPT-3 [3], and T5 [28]. The complexity of these models has been constantly increasing. The size of newer generations of the models has made training computationally prohibitive for all research organizations and is only available to large corporations. Training also requires huge amounts of training data, which do not exist for most languages. Thus, most of these large models have been trained only for a few very well-resourced languages, chiefly English, or in a massively multilingual fashion.

The BERT model was pre-trained on two tasks simultaneously, a masked token prediction and next sentence prediction. For the masked token prediction, 15% of tokens in the training corpus were randomly masked before training. The training dataset was augmented by duplicating the training corpus a few times, with each copy having different randomly selected tokens masked. The next sentence prediction task attempts to predict if two given sentences appear in a natural order.

The RoBERTa [14] model uses the same architecture as BERT, but drops the next sentence prediction task, as it was shown that it does not contribute to the model performance. The masked token prediction task was changed so that the tokens are randomly masked on the fly, i.e. a different subset of tokens is masked in each training epoch.

Both BERT and RoBERTa were released in different sizes. Base models use 12 hidden transformer layers of size 768. Large models use 24 hidden transformer layers of size 1024. Smaller-sized BERT models exist using knowledge distillation from pre-trained larger models [11].

A few massively multilingual models were trained on 100 or more languages simultaneously. Notable released variants are multilingual BERT (mBERT) [5] and XLM-RoBERTa (XLM-R) [4]. While multilingual BERT models perform well for the trained languages, they lag behind the monolingual models [36, 33]. Examples of recently released monolingual BERT models for various languages are Finnish [36], Swedish [16], Estonian [30], Latvian [37], etc.

The Slovene language is supported by the aforementioned massively multilingual models and by the trilingual CroSloEngual BERT model [33], which has been trained on three languages,

Croatian, Slovene, and English. No monolingual transformer model for Slovene has been previously released.

3 SLOBERTA

The presented SloBERTa model is closely related to the French Camembert model [18], which uses the same architecture and training approach as the RoBERTa base model [14], but uses a different tokenization model. In this section, we describe the training datasets, the architecture, and the training procedure of SloBERTa.

3.1 Datasets

Training a successful transformer language model requires a large dataset. We combined five large Slovene corpora in our training dataset. Gigafida 2.0 [13] is a general language corpus, composed of fiction and non-fiction books, newspapers, school textbooks, texts from the internet, etc. The Janes corpus [9] is composed of several subcorpora. Each subcorpus contains texts from a certain social medium or a group of similar media, including Twitter, blog posts, forum conversations, comments under articles on news sites, etc. We used all Janes subcorpora, except Janes-tweet, since the contents of that subcorpus are encoded and need to be individually downloaded from Twitter, which is a lengthy process, as Twitter limits the access speed. KAS (Corpus of Academic Slovene) [8] consists of PhD, MSc, MA, Bsc, and BA theses written in Slovene between 2000 and 2018. SiParl [23] contains minutes of Slovene national assembly between 1990 and 2018. SlWaC [15] is a web corpus collected from the .si top-level web domain. All corpora used are listed in Table 1 along with their sizes.

Table 1: Corpora used in training of SloBERTa with their sizes in billion of tokens and words. Janes* corpus does not include Janes-tweet subcorpus.

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

3.2 Data preprocessing

We deduplicated the corpora, using the Onion tool [26]. We split the deduplicated corpora into three sets, training (99%), validation (0.5%), and test (0.5%). Independently of the three splits, we prepared a smaller dataset, one 15th of the size of the whole dataset, by randomly sampling the sentences. We used this smaller dataset to train a sentencepiece model[1], which is used to tokenize and encode the text into subword byte-pair-encodings (BPE). The sentencepiece model trained for SloBERTa has a vocabulary containing 32,000 subword tokens.

3.3 Architecture and training

SloBERTa has 12 transformer layers, which is equivalent in size to BERT-base and RoBERTa-base models. The size of each transformer layer is 768. We trained the model for 200,000 steps (about 98 epochs) on the Slovene corpora, described in Section 3.1. The model supports the maximum input sequence length of 512 subword tokens.

SloBERTa was trained as a masked language model, using fairseq toolkit [22]. 15% of the input tokens were randomly masked, and the task was to predict the masked tokens. We used the whole-word masking, meaning that if a word was split into more subtokens and one of them was masked, all the other subtokens pertaining to that word were masked as well. Tokens were masked dynamically, i.e. in each epoch, a different subset of tokens were randomly selected to be masked.

4 EVALUATION

We evaluated SloBERTa on five tasks: named-entity recognition (NER), part-of-speech tagging (POS), dependency parsing (DP), sentiment analysis (SA), and word analogy (WA). We used the labeled ssj500k corpus [12, 6] for fine-tuning SloBERTa on each of the NER, POS and DP tasks. For NER, we limited the scope to three types of named entities (person, location, and organization). We report the results as a macro-average F_1 score of these three classes. For POS-tagging, we used UPOS tags, the results are reported as a micro-average F_1 score. For DP, we report the results as a labeled attachement score (LAS). The SA classifier was fine-tuned on a dataset composed of Slovenian tweets [20, 21], labeled as either "positive", "negative", or "neutral". We report the results as a macro-average F_1 score.

Traditional WA task measures the distance between word vectors in a given analogy (e.g., man : king ≈ woman : queen). For contextual embeddings such as BERT, the task has to be modified to make sense. First, word embeddings from transformers are generally not used on their own, rather the model is fine-tuned. Four words from an analogy also do not provide enough context for use with transformers. In our modification, we input the four words of an analogy in a boilerplate sentence "If the word [word1] corresponds to the word [word2], then the word [word3] corresponds to the word [word4]." We then masked [word2] and attempted to predict it using masked token prediction. We used Slovene part of the multilingual culture-independent word analogy dataset [32]. We report the results as an average precision@5 (the proportion of the correct [word2] analogy words among the 5 most probable predictions).

We compared the performance of SloBERTa with three other transformer models supporting Slovene, CroSloEngual BERT (CSE-BERT) [33], multilingual BERT (mBERT) [5], and XLM-RoBERTa (XLM-R) [4]. Where sensible, we also included the results achieved with training a classifier model using Slovene ELMo [31] and fastText embeddings.

We fine-tuned the transformer models on each task by adding a classification head on top of the model. The exception is the DP task, where we used the modified dep2label-bert tool [29, 10]. For ELMo and fastText, we extracted embeddings from the training datasets and used them to train token-level and sentence-level classifiers for each task, except for the DP. The classifiers are composed of a few LSTM layer neural networks. For the DP task, we used the modified SuPar tool, based on the deep biaffine attention [7]. The details of the evaluation process are presented in [34].

The results are shown in Table 2. The results of ELMo and fastText, while comparable between each other, are not fully comparable with the results of transformer models as the classifier training approach is different.

[1] https://github.com/google/sentencepiece

Table 2: Results of Slovene transformer models.

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**

On the NER, POS, SA, and WA tasks, SloBERTa outperforms all other models/embeddings. For the POS-tagging, the differences between the models are small, except for fastText, which performs much worse. ELMo, surprisingly, outperforms all transformer models on the DP task. However, it performs worse on the other tasks. SloBERTa performs worse than CSE-BERT on the DP task, but beats other multilingual models.

The success of ELMo on the DP task can be partially explained by the different tools used for training the classifiers. Further work needs to be done to fully evaluate the difference and success of ELMo embeddings on this task.

The performance on the SA task is limited by the low inter-annotator agreement [20]. The reported average of F_1 scores for positive and negative class is 0.542 for inter-annotator agreement and 0.726 for self-agreement. Using the same measure (average of F_1 for positive and F_1 for negative class), SloBERTa scores 0.667, and mBERT scores 0.593.

On the WA task, most models perform poorly. This is expected because very little context was provided on the input, and the transformer models need a context to perform well. SloBERTa significantly outperforms other models, not only because it was trained only on Slovene data, but largely because its tokenizer is adapted to only Slovene language and does not need to cover other languages.

5 CONCLUSIONS

We present SloBERTa, the first monolingual transformer-based masked language model trained on Slovene texts. We show that SloBERTa large pretrained masked language model outperforms existing comparable multilingual models supporting Slovene on four tasks, NER, POS-tagging, sentiment analysis, and word analogy. The performance on the DP task is competitive, but lags behind some of the existing models.

In further work we intend to compare improvement of BERT-like monolingual models over multilingual models for other languages.

The pre-trained SloBERTa model is publicly available via CLARIN.SI[2] and Huggingface[3] repositories. We make the code, used for preprocessing the corpora and training the SloBERTa, publicly available[4].

ACKNOWLEDGMENTS

The work was partially supported by the Slovenian Research Agency (ARRS) core research programmes P6-0411 and project J6-2581, as well as the Ministry of Culture of Republic of Slovenia through project Development of Slovene in Digital Environment (RSDO). This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

[2]http://hdl.handle.net/11356/1397
[3]https://huggingface.co/EMBEDDIA/sloberta
[4]https://github.com/clarinsi/Slovene-BERT-Tool

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. ArXiv preprint 2108.07258. (2021).

[3] Tom Brown et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. Volume 33, 1877–1901.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. DOI: 10.18653/v1/N19-1423.

[6] Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The universal dependencies treebank for Slovenian. In *Proceeding of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*.

[7] Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of 5th International Conference on Learning Representations, ICLR*.

[8] Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2021. The KAS corpus of Slovenian academic writing. *Language Resources and Evaluation*, 55, 2, 551–583.

[9] Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2016. Janes v0. 4: korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 4, 2, 67–99.

[10] Carlos Gómez-Rodríguez, Michalina Strzyz, and David Vilares. 2020. A unifying theory of transition-based and sequence labeling parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3776–3793. DOI: 10.18653/v1/2020.coling-main.336.

[11] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. (2020). arXiv: 1909.10351 [cs.CL].

[12] Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2019. Training corpus ssj500k 2.2. Slovenian language resource repository CLARIN.SI. (2019).

[13] Simon Krek, Tomaž Erjavec, Andraž Repar, Jaka Čibej, Spela Arhar, Polona Gantar, Iztok Kosem, Marko Robnik, Nikola Ljubešić, Kaja Dobrovoljc, Cyprian Laskowski, Miha Grčar, Peter Holozan, Simon Šuster, Vojko Gorjanc, Marko Stabej, and Nataša Logar. 2019. Gigafida 2.0: Korpus pisne standardne slovenščine. viri.cjvt.si/gigafida. (2019).

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv preprint 1907.11692. (2019).

[15] Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *International Conference on Text, Speech and Dialogue*. Springer, 395–402.

[16] Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. ArXiv preprint 2007.01658. (2020).

[17] Gary Marcus and Ernest Davis. 2021. Has AI found a new foundation? *The Gradient*. 11 September 2021.

[18] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219.

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint 1301.3781*.

[20] Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter sentiment classification: the role of human annotators. *PLOS ONE*, 11, 5.

[21] Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Twitter sentiment for 15 european languages. Slovenian language resource repository CLARIN.SI. (2016). http://hdl.handle.net/11356/1054.

[22] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: a fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

[23] Andrej Pančur and Tomaž Erjavec. 2020. The siParl corpus of Slovene parliamentary proceedings. In *Proceedings of the Second ParlaCLARIN Workshop*, 28–34.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing EMNLP*, 1532–1543.

[25] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. DOI: 10.18653/v1/N18-1202.

[26] Jan Pomikálek. 2011. *Removing boilerplate and duplicate content from web corpora*. PhD thesis. Masaryk university, Brno, Czech Republic.

[27] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog. 2019 Feb 24. (2019).

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21, 1–67.

[29] Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Viable dependency parsing as sequence labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 717–723. DOI: 10.18653/v1/N19-1077.

[30] Hasan Tanvir, Claudia Kittask, and Kairit Sirts. 2020. EstBERT: A pretrained language-specific BERT for Estonian. arXiv preprint 2011.04784. (2020).

[31] Matej Ulčar and Marko Robnik-Šikonja. 2020. High quality ELMo embeddings for seven less-resourced languages. In *Proceedings of the 12th Language Resources and Evaluation Conference, LREC 2020*, 4733–4740.

[32] Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4067–4073.

[33] Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Proceedings of Text, Speech, and Dialogue, TSD 2020*, 104–111.

[34] Matej Ulčar, Aleš Žagar, Carlos S. Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja. 2021. Evaluation of contextual embeddings on less-resourced languages. ArXiv preprint 2107.10614. (2021).

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

[36] Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076. (2019).

[37] Artūrs Znotiņš and Guntis Barzdiņš. 2020. LVBERT: Transformer-based model for Latvian language understanding. In *Human Language Technologies–The Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020*. Volume 328, 111.