This paper was published in **TSD 2021** – please cite the published version https://doi.org/10.1007/978-3-030-83527-9_17 instead.

RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model

Milan Straka[0000−0003−3295−5576], Jakub Náplava[0000−0003−2259−1377], Jana Straková[0000−0003−0075−2408], and David Samuel[0000−0003−2866−1022]

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské nám. 25, 118 00 Prague, Czech Republic
{straka,naplava,strakova,samuel}@ufal.mff.cuni.cz

Abstract. We present RobeCzech, a monolingual RoBERTa language representation model trained on Czech data. RoBERTa is a robustly optimized Transformer-based pretraining approach. We show that RobeCzech considerably outperforms equally-sized multilingual and Czech-trained contextualized language representation models, surpasses current state of the art in all five evaluated NLP tasks and reaches state-of-the-art results in four of them. The RobeCzech model is released publicly at https://hdl.handle.net/11234/1-3691 and https://huggingface.co/ufal/robeczech-base.

Keywords: RobeCzech · Czech RoBERTa · RoBERTa

1 Introduction

We introduce RobeCzech: Czech RoBERTa, a Czech contextualized language representation model based on the Transformer architecture and trained solely on Czech data. RobeCzech is a monolingual version of RoBERTa [23], a robustly optimized BERT [8] pretraining approach.

In this paper, we describe the RobeCzech training process and we evaluate RobeCzech in comparison with current multilingual and Czech-trained contextualized language representation models: multilingual BERT [8], multilingual XLM-RoBERTa [6] (base and large), Slavic BERT [1] tuned on 4 Slavic languages, including Czech; and Czert [36], another monolingual, Czech BERT model.

We show that RobeCzech considerably outperforms all models of similar size, and at the same time, it reaches new state-of-the-art results in four NLP tasks: morphological tagging and lemmatization, dependency parsing, named entity recognition and semantic parsing. In the last evaluated task, the sentiment analysis, RobeCzech also improves over state of the art and delivers the best results of all models of similar size, only being surpassed by XLM-RoBERTa large [6], a model 4 times the size of all the other evaluated models (Table 1).

We release the RobeCzech model for public use.

2 Related Work

Contextualized language representation models have recently accelerated progress in NLP. Significant advances have been reached particularly with Bidirectional Encoder Representations from Transformers, widely known as BERT [8], inspiring interest in Transformer-like architectures. We especially highlight RoBERTa [23] and its derivation XLM-RoBERTa [6].

The above mentioned language representation models were trained either only on English or as multilingual, though with an (implicit) strength in the most represented languages (i. e., English). Therefore, research has recently been focusing on monolingual BERT models, giving birth to national BERT mutations, e.g. French [26], Finnish [43], Romanian [27] and Czech [36].

Our model is similar to the above mentioned Czert [36] in the sense that it is also a Czech contextualized language representation model, but unlike Czert, which is based on BERT, we trained a Czech version of RoBERTa. According to both the original Czert results [36] and the hereby presented evaluation on five NLP tasks, RobeCzech is better than Czert in all experiments by a considerable margin.

3 Training the Czech RoBERTa

We trained RobeCzech on a collection of the following publicly available texts:

- SYN v4 [21], a large corpus of contemporary written Czech, 4,188M tokens;
- Czes [7], a collection of Czech newspaper and magazine articles, 432M tokens;
- documents with at least 400 tokens from the Czech part of the web corpus W2C [24,25], tokenized with MorphoDiTa [40], 16M tokens;
- plain texts extracted from Czech Wikipedia dump 20201020 using WikiExtractor,[1] tokenized with MorphoDiTa [40], 123M tokens.

All these corpora contain whole documents, even if the SYN v4 is block-shuffled (blocks with at most 100 words respecting sentence boundaries are permuted in a document) and in total contain 4,917M tokens.

The texts are tokenized into subwords with a byte-level BPE (BBPE) tokenizer [33]. The tokenizer is trained on the entire corpus and we limit its vocabulary size to 52,000 items.

The RobeCzech model is trained using the official code released in the Fairseq library.[2] The training batch size is 8,192 and each training batch consists of sentences sampled contiguously, even across document boundaries, such that the total length of each sample is at most 512 tokens (*FULL-SENTENCES* setting [23]). We use Adam optimizer [16] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ to minimize the masked language-modeling objective. The learning rate is adapted using the polynomial decay schema with 10,000 warmup updates and the peak learning rate set to $7 \cdot 10^{-4}$. A total amount of 91,075 optimization steps were performed, which took approximately 3 months on 8 QUADRO P5000 GPU cards.

[1] https://github.com/attardi/wikiextractor

[2] https://github.com/pytorch/fairseq/blob/master/examples/roberta/

Table 1. Number of parameters.

Model	**Embedding**	**Transformer**	**Total Parameters**
mBERT uncased	82M	85M	167M
Czert	24M	85M	109M
Slavic BERT	92M	85M	177M
XLM-R base	192M	85M	277M
XLM-R large	257M	302M	559M
RobeCzech	40M	85M	125M

4 Evaluation Tasks

We evaluate our Czech RoBERTa model on five NLP tasks in comparison with a variety of recently proposed mono- and multi-lingual contextualized language representation models (to our best knowledge, these are all publicly available models trained at least partially on Czech):

- **mBERT [8]:** well-known multilingual BERT language representation model.
- **Czert [36]:** the first Czech monolingual model based on BERT.
- **Slavic BERT [1]:** multilingual BERT tuned specifically for NER on 4 Slavic languages data (Russian, Bulgarian, Czech and Polish).
- **XLM-RoBERTa [6], base and large:** multilingual contextualized representations trained at large scale.

Except for XLM-RoBERTa large, which is 4 times larger than others, all models are of *base* size [8], see Table 1.

We evaluate RobeCzech in five NLP tasks, three of them leveraging frozen contextualized word embeddings, two approached with fine-tuning:

- **morphological analysis and lemmatization:** frozen contextualized word embeddings,
- **dependency parsing**: frozen contextualized word embeddings,
- **named entity recognition:** frozen contextualized word embeddings,
- **semantic parsing:** fine-tuned,
- **sentiment analysis:** fine-tuned.

4.1 Morphological Tagging and Lemmatization on PDT 3.5

Dataset We evaluate the morphological POS tagging and lemmatization on the morphological layer of the *Prague Dependency Treebank 3.5* [14].

Metric The morphological POS tagging and lemmatization is evaluated using accuracy.

Architecture We adopt the *UDPipe 2* architecture [38], reproducing the methodology of [39]. After embedding input words, three bidirectional LSTM layers [15] are applied, followed by a softmax classification layer for POS and lemmas. In case of lemmas, the network predicts a simple edit script from input form to desired lemma. Since edit patterns are shared between lemmas due to regularities in morphology, the output categorization layer is reduced from the full vocabulary to only 1568 classes (in PDT 3.5 [14]). In all our experiments, we use the same word embeddings as [39]: pretrained `word2vec` embeddings [28], end-to-end word embeddings and character-level word embeddings [5,11,22]. The contextualized word embeddings are used frozen-style as additional inputs to the neural network.

4.2 Dependency Parsing on PDT 3.5

Dataset We evaluate the dependency parsing on the analytical layer of the *Prague Dependency Treebank 3.5* [14].

Metric In evaluation, we compute both the unlabeled attachment score (UAS) and labeled attachment score (LAS).

Architecture We perform dependency parsing jointly with POS tagging and lemmatization, following the experiments of [39] showing that this approach is superior to using predicted POS tags and lemmas on input. We utilize the *UDPipe 2* architecture [39]: after embeddings input words and three bidirectional LSTM layers [15], a biaffine attention layer [10] produces labeled dependency trees. The input word embeddings are the same as in the previous Section 4.1 and the contextualized word embeddings are additionally concatenated to the baseline input.

4.3 Morphosyntactic Analysis on Universal Dependencies

Dataset We further evaluate the joint morphosyntactic analysis on the *UD Czech PDT* treebank of the *Universal Dependencies 2.3* [30].

Metric We use the standard evaluation script from *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* [45], which produces the following metrics:

- **UPOS** – universal POS tags accuracy,
- **XPOS** – language-specific POS tags accuracy,
- **UFeats** – universal subset of morphological features accuracy,
- **Lemmas** – lemmatization accuracy,
- **UAS** – unlabeled attachment score, **LAS** – labeled attachment score,
- **MLAS** – morphology-aware LAS, **BLEX** – bi-lexical dependency score.

Architecture Following Section 4.2, we employ the *UDPipe 2* [39] architecture with frozen contextualized word embeddings.

4.4 Named Entity Recognition

Dataset We evaluate the Czech NER on all versions of the *Czech Named Entity Corpus*, both the original [35] with nested entities and the CoNLL version [19] with reduction to flat entities only.

Metric The standard evaluation metric for NER is F1 score computed over detected named entities spans.

Architecture We reproduce the the current NER SoTA architecture [41], using the *LSTM-CRF* and *seq2seq* variants for flat and nested NER, respectively. All experiments include the Czech FastText word embeddings [4] of dimension 300, end-to-end trained word embeddings and character-level word embeddings [5,11,22] as inputs to the network. The contextualized word embeddings are used as frozen, additional inputs to the network.

4.5 Semantic Parsing on Prague Tectogrammatical Graphs

Dataset We use the *Prague Tectogrammatical Graphs* (PTG) provided for the CoNLL 2020 shared task, *Cross-Framework Meaning Representation Parsing* (MRP 2020) [31]. The original annotation comes from the tectogrammatical layer of the *Prague Dependency Treebank* [14]; the graphs for the shared task were obtained by relaxing its original limitation to trees – for example by explicitly modeling co-reference by additional edges instead of special node attributes [44].

Metric We employ the official metric from MRP 2020, which first finds the maximum common edge subgraph to align the evaluated and the target graph. Then, it computes the micro-averaged F1 score over different features of the semantic graphs – top nodes, node labels, node properties, anchors, edges between nodes and edge attributes.

Architecture We reimplement the current SoTA architecture for PTG parsing called *PERIN* [34]. This model does not assume any hard-coded ordering of the graph nodes, but instead dynamically finds the best matching between the predicted and the target ones.

Following *UDify* [17], we compute the contextualized subword embedding by taking the weighted sum of all hidden layers in a language representation model. The scalar weight for each layer is a learnable parameter. To obtain a single embedding for every token, we sum the embeddings of all its subwords. Finally, the summed embeddings are normalized with layer normalization [3] to stabilize the training.

The pretrained encoder is finetuned with a lower learning rate than the rest of the model. The learning rate follows the inverse square root schedule with warmup and is frozen for the first 2000 steps before the warmup starts. The warmup phase takes 6000 steps and the learning rate peak is $6 \cdot 10^{-5}$.

4.6 Sentiment Analysis

Dataset We evaluate sentiment analysis on Czech Facebook dataset (CFD) [12,13]. This dataset contains 2,587 positive, 5,174 neutral and 1,991 negative posts (the 248 bipolar posts are ignored, following [13,36]).

Metric The performance is evaluated using macro-averaged F1 score. Because the dataset has no designed test set, we follow the approach of the dataset authors [13] and perform 10-fold cross-validation, reporting mean and standard deviation of the folds' F1 scores.

Table 2. Overall results.

	Morphosynt.		**Morphosynt.**		**NER**		**Semant.**	**Sentim.**
	PDT3.5		**UD2.3**		**CNEC1.1**		**PTG**	**CFD**
	POS	**LAS**	**XPOS**	**LAS**	**nested**	**flat**	**Avg.**	**F1**
mBERT	98.00	89.74	97.61	92.34	86.71	86.45	90.62	75.43
Czert	98.43	90.68	98.07	93.13	85.38	84.69	90.66	78.52
Slavic BERT	97.70	88.50	97.29	91.49	85.85	85.12	91.27	74.85
XLM-R base	97.62	88.14	97.29	91.30	83.25	82.76	91.55	79.40
XLM-R large	98.41	91.27	98.15	93.49	87.41	86.86	92.11	**82.29**
RobeCzech	**98.50**	**91.42**	**98.31**	**93.77**	**87.82**	**87.47**	**92.36**	80.13
previous SoTA	98.05	89.89	97.71	93.38	86.88	86.57	92.24	76.55

Architecture We employ the standard text classification architecture consisting of a BERT encoder, followed by a softmax-activated classification layer processing the computed embedding of the given document text obtained from the `CLS` token embedding from the last layer [8,23].

We train the models using a lazy variant of the Adam optimizer [16] with a batch size of 64. During the first epoch, the BERT encoder is frozen and only the classifier is trained with the default learning rate of 10^{-3}. From the second epoch, the whole model is updated, starting by 4 epochs of cosine warm-up from zero to a specified peak learning rate, followed by 10 epochs of cosine decay back to zero.

We consider peak learning rates $10^{-5}, 2 \cdot 10^{-5}, 3 \cdot 10^{-5}$ and $5 \cdot 10^{-5}$. In order to choose the peak learning rate, we put aside random 10% of the train data for each fold as a development set and evaluate each trained model on its corresponding development set. Finally, we choose a single peak learning rate for every model according to the 10-fold means of the development macro-averaged F1 scores. The selected peak learning rates are reported for each evaluated model.

5 Results

Table 2 summarizes the overall results of all considered language representation models in all evaluated tasks. RobeCzech improves over current state of the art in all five evaluated NLP tasks, and at the same time, clearly outperforms current multilingual and Czech-trained contextualized language representation models, being surpassed only in one of the five tasks by a model 4 times its size (XLM-RoBERTa large [6], Table 2). Notably, RobeCzech reaches 25% error reduction in POS tagging both on PDT 3.5 and UD 2.3, and 15% error reduction in dependency parsing on PDT 3.5, significantly improving performance of Czech morphosyntactic analysis.

Furthermore, for each of the evaluated tasks, we show the detailed results in Tables 3, 4, 5, 6, 7 and 8.

The results demonstrate that the large variant of XLM-RoBERTa reaches considerably better results compared to base size of other multilingual models.

Table 3. Morpological tagging and lmmatization on PDT3.5.

Model	**Without Dictionary**			**With Dictionary**		
	POS	**Lemmas**	**Both**	**POS**	**Lemmas**	**Both**
mBERT	97.86	98.69	97.21	98.00	98.96	97.59
Czert	98.30	98.73	97.65	98.43	98.98	98.02
Slavic BERT	97.51	98.58	96.81	97.70	98.89	97.27
XLM-R base	97.43	98.56	96.76	97.62	98.85	97.20
XLM-R large	98.30	98.76	97.69	98.41	98.98	98.01
RobeCzech	**98.43**	**98.79**	**97.83**	**98.50**	**99.00**	**98.11**
Morče (2009) [37]	—	—	—	95.67	—	—
MorphoDiTa (2016) [40]	—	—	—	95.55	97.85	95.06
LemmaTag (2018) [18]	96.90	98.37	—	—	—	—
UDPipe 2+mBERT+Flair [39]	97.94	98.75	97.31	98.05	98.98	97.65

Table 4. Dependency parsing on PDT3.5.

Model	**UAS**	**LAS**	***Joint POS***	***Joint Lemmas***
mBERT	93.01	89.74	*97.62*	*98.49*
Czert	93.57	90.68	*98.10*	*98.53*
Slavic BERT	92.14	88.50	*97.20*	*98.29*
XLM-R base	91.80	88.14	*97.22*	*98.34*
XLM-R large	94.07	91.27	*98.12*	*98.54*
RobeCzech	**94.14**	**91.42**	***98.28***	***98.62***
UDPipe 2+mBERT+Flair [39]	93.07	89.89	*97.72*	*98.51*

Table 5. Morphosyntactic analysis on UD 2.3. Models marked ^{f} are fine-tuned, otherwise with frozen embeddings.

Model	**UPOS**	**XPOS**	**UFeats**	**Lemmas**	**UAS**	**LAS**	**MLAS**	**BLEX**
mBERT	99.31	97.61	97.55	99.06	94.27	92.34	87.75	89.91
Czert	99.32	98.07	98.05	99.09	94.75	93.13	89.19	90.92
Slavic BERT	99.22	97.29	97.22	98.99	93.53	91.49	86.37	88.79
XLM-R base	99.18	97.29	97.24	99.02	93.32	91.30	86.18	88.62
XLM-R large	99.36	98.15	98.10	99.17	95.15	93.49	89.64	91.40
RobeCzech	**99.36**	**98.31**	**98.28**	**99.18**	**95.36**	**93.77**	**90.18**	**91.82**
UDPipe 2 +mBERT+Flair [39]	99.34	97.71	97.67	99.12	94.43	92.56	88.09	90.22
UDify^f [17]	99.24	—	94.77	98.93	95.07	93.38	—	—
Czert^f [36]	99.30	—	—	—	—	—	—	—

Table 6. Named entity recognition F1 scores (3 runs average) in comparison with previous reports. Models marked ^{f} are fine-tuned, otherwise with frozen embeddings.

Model	**CNEC1.1**	**CNEC2.0**	**CoNLL** **CNEC1.1**	**CoNLL** **CNEC2.0**
mBERT	86.71	84.21	86.45	87.04
Czert	85.38	82.84	84.69	85.33
Slavic BERT	85.85	82.71	85.12	85.28
XLM-R base	83.25	80.33	82.76	82.85
XLM-R large	87.41	84.46	86.86	87.06
RobeCzech	**87.82**	**85.51**	**87.47**	**87.49**
seq2seq+mBERT [39,41]	86.73	84.66	—	—
seq2seq+mBERT+Flair [39,41]	86.88	84.27	—	—
LSTM-CRF, LDA [20]	—	—	81.77	—
LSTM-CRF [42]	83.15	—	83.27	84.22
LSTM-CRF+BERT [29]	—	—	—	86.39
Czert^f [36]	—	—	86.27	—
mBERT^f [8], by [36]	—	—	86.23	—
Slavic BERT^f [1], by [36]	—	—	86.57	—

Table 7. Semantic parsing F1 scores on Prague Tectogrammatical Graphs.

Model	**Labels**	**Properties**	**Anchors**	**Edges**	**Attributes**	**Average**
mBERT	95.72	92.60	97.20	80.77	72.83	90.62
Czert	95.72	92.69	97.23	80.91	72.37	90.66
Slavic BERT	95.92	92.91	97.51	82.48	75.08	91.27
XLM-R base	96.09	93.12	97.60	83.03	76.16	91.55
XLM-R large	96.42	93.31	97.92	84.46	77.89	92.11
RobeCzech frozen	95.85	92.76	97.41	82.60	74.95	91.23
RobeCzech	**96.57**	**93.58**	**97.97**	**84.92**	**78.29**	**92.36**
HUJI-KU+mBERT [2]	—	72.44	72.10	44.91	—	58.49
HIT-SCIR+mBERT [9]	84.14	79.01	92.34	64.96	47.68	77.93
Hitachi+mBERT [32]	87.69	91.48	93.99	76.90	66.07	87.35
ÚFAL+XLM-R large [34]	96.23	93.56	97.86	84.61	78.62	92.24

Table 8. Sentiment analysis 10-fold macro F1 scores on Czech Facebook dataset.

Model	**10-fold Macro F1**	**10-fold Std**	**Chosen LR**
mBERT	75.43	±1.38	$5 \cdot 10^{-5}$
Czert	78.52	±1.16	$2 \cdot 10^{-5}$
Slavic BERT	74.85	±1.27	$5 \cdot 10^{-5}$
XLM-R base	79.40	±1.07	$1 \cdot 10^{-5}$
XLM-R large	**82.29**	±1.19	$1 \cdot 10^{-5}$
RobeCzech	80.13	±1.21	$3 \cdot 10^{-5}$
Czert [36]	76.55	—	$3 \cdot 10^{-6}$
MaxEnt [13]	69.4	—	—

Yet, RobeCzech still surpasses it on four tasks, most notably in the frozen scenario. We hypothesize that in the frozen scenario the larger model cannot capitalize on its superior capacity, compared to for example sentiment analysis, where its capacity proves determining.

6 Conclusion

We introduced RobeCzech, a Czech contextualized language representation model based on RoBERTa. We described the training process and we evaluated RobeCzech in comparison with, to our best knowledge, currently known multilingual and Czech-trained contextualized language representation models. We show that RobeCzech considerably improves over state of the art in all five evaluated NLP tasks. Notably, it yields 25% error reduction in POS tagging both on PDT 3.5 and UD 2.3 and 15% error reduction in dependency parsing on PDT 3.5. We publish RobeCzech publicly at https://hdl.handle.net/11234/1-3691 and https://huggingface.co/ufal/robeczech-base.

Acknowledgements

The research described herein has been supported by the Czech Science Foundation grant No. GX20-16819X, Mellon Foundation Award No. G-1901-06505 and by the Grant Agency of Charles University 578218. The resources used have been provided by the LINDAT/CLARIAH-CZ Research Infrastructure, project No. LM2018101 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. pp. 89–93. Association for Computational Linguistics, Florence, Italy (Aug 2019)
2. Arviv, O., Cui, R., Hershcovich, D.: HUJI-KU at MRP 2020: Two Transition-based Neural Parsers. In: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing. pp. 73–82. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.conll-shared.7
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
5. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. CoRR (2014)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020)

7. Czes (2011), http://hdl.handle.net/11858/00-097C-0000-0001-CCCF-C, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186 (2019)
9. Dou, L., Feng, Y., Ji, Y., Che, W., Liu, T.: HIT-SCIR at MRP 2020: Transition-based Parser and Iterative Inference Parser. In: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing. pp. 65–72. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.conll-shared.6
10. Dozat, T., Manning, C.D.: Deep Biaffine Attention for Neural Dependency Parsing. CoRR **abs/1611.01734** (2016)
11. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks pp. 5–6 (2005)
12. Habernal, I., Ptáček, T., Steinberger, J.: Facebook data for sentiment analysis (2013), http://hdl.handle.net/11858/00-097C-0000-0022-FE82-7, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
13. Habernal, I., Ptáček, T., Steinberger, J.: Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 65–74. Association for Computational Linguistics, Atlanta, Georgia (Jun 2013)
14. Hajič, J., et al.: Prague Dependency Treebank 3.5 (2018), http://hdl.handle.net/11234/1-2621, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
15. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. **9**(8), 1735–1780 (November 1997)
16. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (12 2014)
17. Kondratyuk, D., Straka, M.: 75 Languages, 1 Model: Parsing Universal Dependencies Universally. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2779–2795. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1279
18. Kondratyuk, D., Gavenčiak, T., Straka, M., Hajič, J.: LemmaTag: Jointly Tagging and Lemmatizing for Morphologically Rich Languages with BRNNs. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4921–4928. Association for Computational Linguistics (2018)
19. Konkol, M., Konopík, M.: CRF-Based Czech Named Entity Recognizer and Consolidation of Czech NER Research. pp. 153–160 (09 2013)
20. Konopík, M., Prazák, O.: LDA in Character-LSTM-CRF Named Entity Recognition. In: Sojka, P., Horák, A., Kopecek, I., Pala, K. (eds.) Text, Speech, and Dialogue - 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings. Lecture Notes in Computer Science, vol. 11107, pp. 58–66. Springer (2018)
21. Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal,

M., Truneček, P., Vondřička, P., Zasina, A.: SYN v4: large corpus of written Czech (2016), http://hdl.handle.net/11234/1-1846, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

22. Ling, W., Luís, T., Marujo, L., Astudillo, R.F., Amir, S., Dyer, C., Black, A.W., Trancoso, I.: Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. CoRR (2015)
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR **abs/1907.11692** (2019)
24. Majliš, M.: W2C – Web to Corpus – Corpora (2011), http://hdl.handle.net/11858/00-097C-0000-0022-6133-9, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
25. Majliš, M., Žabokrtský, Z.: Language richness of the web. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 2927–2934. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)
26. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a Tasty French Language Model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020)
27. Masala, M., Ruseti, S., Dascalu, M.: RoBERT – A Romanian BERT Model. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6626–6637. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020)
28. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc. (2013)
29. Štěpán Müller: Text Summarization Using Named Entity Recognition. Master's thesis, Czech Technical University in Prague (2020)
30. Nivre, J., et al.: Universal Dependencies 2.3 (2018), http://hdl.handle.net/11234/1-2895, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
31. Oepen, S., Abend, O., Abzianidze, L., Bos, J., Hajic, J., Hershcovich, D., Li, B., O'Gorman, T., Xue, N., Zeman, D.: MRP 2020: The Second Shared Task on Cross-Framework and Cross-Lingual Meaning Representation Parsing. In: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing. pp. 1–22. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.conll-shared.1
32. Ozaki, H., Morio, G., Koreeda, Y., Morishita, T., Miyoshi, T.: Hitachi at MRP 2020: Text-to-Graph-Notation Transducer. In: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing. pp. 40–52. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.conll-shared.4
33. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
34. Samuel, D., Straka, M.: ÚFAL at MRP 2020: Permutation-invariant Semantic Parsing in PERIN. In: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework

Meaning Representation Parsing. pp. 53–64. Association for Computational Linguistics, Online (Nov 2020)

35. Ševčíková, M., Žabokrtský, Z., Krůza, O.: Named Entities in Czech: Annotating Data and Developing NE Tagger. In: Matoušek, V., Mautner, P. (eds.) Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue. Lecture Notes in Computer Science, vol. 4629, pp. 188–195. Springer, Berlin / Heidelberg (2007)
36. Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., Konopík, M.: Czert – Czech BERT-like Model for Language Representation (2021)
37. Spoustová, D.j., Hajič, J., Raab, J., Spousta, M.: Semi-Supervised Training for the Averaged Perceptron POS Tagger. In: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). pp. 763–771. Association for Computational Linguistics (Mar 2009)
38. Straka, M.: UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning. pp. 197–207. Association for Computational Linguistics, Stroudsburg, PA, USA (2018)
39. Straka, M., Straková, J., Hajič, J.: Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. In: Proceedings of the 22nd International Conference on Text, Speech and Dialogue - TSD 2019. pp. 137–150. Springer International Publishing, Cham / Heidelberg / New York / Dordrecht / London (2019)
40. Straková, J., Straka, M., Hajič, J.: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 13–18. Johns Hopkins University, Baltimore, MD, USA, Association for Computational Linguistics (2014)
41. Straková, J., Straka, M., Hajič, J.: Neural Architectures for Nested NER through Linearization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5326–5331. Association for Computational Linguistics, Stroudsburg, PA, USA (2019)
42. Straková, J., Straka, M., Hajič, J., Popel, M.: Hluboké učení v automatické analýze českého textu. Slovo a slovesnost **80**(4), 306–327 (2019)
43. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: BERT for Finnish. CoRR **abs/1912.07076** (2019)
44. Zeman, D., Hajic, J.: FGD at MRP 2020: Prague Tectogrammatical Graphs. In: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing. pp. 33–39. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.conll-shared.3
45. Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., Petrov, S.: CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 1–21. Association for Computational Linguistics, Brussels, Belgium (Oct 2018)