S. Krajči (ed.): ITAT 2018 Proceedings, pp. 138–143
CEUR Workshop Proceedings Vol. 2203, ISSN 1613-0073, ITAT

Solving Three Czech NLP Tasks End-to-End with Neural Models

Jindřich Libovický, Rudolf Rosa, Jindřich Helcl, and Martin Popel

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University,
Malostranské náměstí 25, 118 00 Praha, Czech republic
surname@ufal.mff.cuni.cz

Abstract: In this work, we focus on three different NLP tasks: image captioning, machine translation, and sentiment analysis. We reimplement successful approaches of other authors and adapt them to the Czech language. We provide end-to-end architectures that achieve state-of-the-art results on all of the tasks within a single sequence learning toolkit. The trained models are available both for download as well as in an online demo.

1 End-to-End Training

Traditionally, solving tasks such as machine translation or sentiment analysis required complex processing pipelines consisting of tools which transformed one explicit representation of the data into another, with the structure of the internal representations defined by the system designer. In machine translation, we would devise explicit word alignment links, extract phrase tables, train a language model, etc.; in sentiment analysis, we could label the data with part-of-speech tags, decode their syntactic structure, and/or assign them with semantic labels. All of these more-or-less linguistically motivated internal representations are not inherently required to produce the desired output, but have been devised as clever and useful ways to break down the large and hard task into smaller and manageable substeps.

With the advent of end-to-end training of deep neural networks (DNN), the need for most of this has been eliminated. In the end-to-end learning paradigm, there is only one model, directly trained to produce the desired outputs from the inputs, without any explicit intermediate representations. The system designer now only has to design a rather generic architecture of the system. It mostly does not enforce any complex explicit representations and processing steps, but rather offers opportunities for the DNN to devise its own notion of intermediate representations and processing steps through training.

This also means that similar architectures can be used to solve very different tasks. Rather than by the nature of the task itself, the structure of the DNN to use is mostly determined by the structure of the input and output – e.g. image inputs are processed by two-dimensional convolutions, while text inputs are processed by one-dimensional convolutions, recurrent units, and/or attentions; classification can produce its output in one step, while text generation is better done iteratively using recurrent decoders; etc.

Thanks to that, a single general framework can be used to solve many different tasks. One just needs to transform the inputs and outputs into a suitable format, define an adequate network structure, and let the system train for a few weeks.

Sadly, the burden of hyperparameter tuning has not been alleviated by DNNs, but rather made worse by the computational costliness of the training. However, with a bit of experience, one is often able to propose a suitable architecture and hyperparameter values at the first attempt, already achieving very competitive results even without any further tuning.

1.1 Our contribution

Most of the papers in the field only evaluate their setups on English datasets. In our work, we try to rectify this shortcoming by reimplementing existing state-of-the-art approaches in the Neural Monkey framework [14] and training them on existing Czech datasets.

Neural Monkey is an open-source toolkit for sequence-to-sequence learning, implemented in the TensorFlow library [1]. The toolkit is designed to be easily extensible in order to support fast prototyping of architectures for various NLP tasks. It is freely available on GitHub[1] under the BSD license, allowing both non-commercial and commercial use of the toolkit.

We decided to focus on three rather varied tasks – sentiment analysis, machine translation, and image captioning. For each of the tasks, we reimplemented one or more existing state-of-the-art architectures within Neural Monkey and trained it on available datasets. Our evaluations show that we manage to reach or surpass state-of-the-art results for all the three tasks.

As we wish to encourage other NLP researchers to focus on Czech language, we make sure that our source codes, our configuration files and our trained models are all freely available to anyone interested to use them, to study them and to build upon them. With our work, we hope to establish well-performing approaches for Czech NLP, as well as to allow e.g. investigation of the internals of the trained models to try to decipher in what ways language seems to be implicitly captured in them. Moreover, we have created a simple web-based demo that allows anyone to easily apply our models to any input data, intended to popularize deep learning and its applications for the Czech audience.

[1] https://github.com/ufal/neuralmonkey

2 Sentiment Analysis

The goal of the sentiment analysis tasks is to decide whether a text expresses positive, neutral or negative judgment on its topic, sometimes also a degree of the positivity or negativity.

2.1 Architecture

We use a state-of-the-art architecture by Lin et al. [20]. The architecture processes the text with a bi-directional LSTM network [15, 11]. After that the attention mechanism is applied several times, each time with a different trained query vector; this is usually referred to as the architecture featuring multiple attention heads. This gives us a set of context vector, each of them being a different weighted average of the LSTM states.

For English, Lin at el. [20] achieved new state-of-the-art results on the Yelp dataset[2] which contains texts of restaurant reviews and the number of stars the users assigned to the review. The goal of the prediction is an automatic assignment of the stars.

After replicating the results on the English dataset, we evaluated the same approach on a Czech dataset. We also experimented with architectures based on processing the input with a convolutional network or a recurrent network followed by max-pooling in time [16].

All models use embeddings of size 300 and a classifier with 100 hidden units. In the experiments with CNN, we used kernels of size 3, 4 and 5 with output dimension 100. The LSTM network used 300 hidden units in both directions. The self-attentive layer used 10 heads and a hidden layer of 300 dimensions.

2.2 Dataset

The largest existing Czech dataset for sentiment analysis is the *CSFD CZ* dataset [12], which is available online[3] under the CC-BY-NC-SA license. It consists of 91,379 movie reviews from ČSFD,[4] a Czechoslovak film database.

The textual reviews are on average 60 tokens long, and bear a rating of 0 to 6 stars, which the authors of the dataset mapped into three classes: negative (0-2 stars), neutral (3-4 stars), and positive (5-6 stars). The three classes are represented rather uniformly, each being assigned to 32%-34% reviews.

We split off 2,000 reviews for validation and another 2,000 for testing (there is no official split of the dataset), retaining the nearly uniform distribution of the classes, as well as other characteristics of the dataset such as average review length.

We use tokenization from the Moses MT toolkit [18] and post-process the tokenization in order to normalize emoticons and repetitive vowels that are often used for emphasis. We use vocabulary of 50k tokens appearing at least 5 times in the training data.

[2] https://www.yelp.com/dataset/

[3] http://liks.fav.zcu.cz/sentiment/

[4] https://www.csfd.cz/

Setup	Accuracy
Most frequent class	35.70 %
Maxpool on embeddings	$80.3 \pm .1$ %
CNN + maxpool	$79.2 \pm .1$ %
SAN on embeddings	$80.1 \pm .1$ %
SAN on LSTM	$80.8 \pm .1$ %
Lenc+ [19]	71.00 %
Brychcín+ [7]	81.53 %

Table 1: Quantitative evaluation of sentiment analysis

2.3 Evaluation

The evaluation in Table 1 show that no matter which particular architecture we use, we achieve accuracies around 81 %. We hypothesize that this already approaches the highest accuracy practically achievable on the dataset, and that all of the model architectures are sufficiently powerful to achieve this accuracy.

Similarly to our approach, Lenc and Hercig [19] experiment with convolutional networks and max-pooling [16], however due to a small vocabulary and limited input length, they report scores which are ten percentage points smaller than ours.

To the best of our knowledge, the best result on this dataset has been reported for the "ME + sspace + Dir" setup of Brychcín+ [7]; they report a ± 0.3 confidence interval for their accuracy, which our best result also falls into. The authors use a complex setup combining a Maximum Entropy classifier with an unsupervised extension that incorporates global context into the classification, based on the assumption that reviews for the same target (movie) tend to bear similar labels; this extension brings them approximately +3 accuracy points. We do not incorporate this mechanism into our setup; in fact, our system does not use the information about the identity of the movie at all. This shows that our model is stronger in a restricted variant of the task – predicting the sentiment solely from the plain text.

3 Machine Translation

Machine translation (MT) is one of the most well-studied problems from NLP. In general, the goal of MT is given a sentence in a source language, generate a sentence in a target language which as similar meaning as possible to the source sentence.

3.1 Architecture

We use our implementation of the self-attentive architecture called the Tranformer [27]. Our implementation

is compatible with the official implementation in Tensor2Tensor [26] and we can thus take advantage of highly optimized training procedure.

The architecture uses the encoder-decoder scheme [3]. Unlike the original sequence-to-sequence models which were based on recurrent neural networks, the Transformer model uses a stack of self-attentive and feed-forward layers.

In the self-attentive layers, we use the state as a query to an attention over the remaining states of the layer and output a weighted combination of the states. This is always followed by a feed-forward layer. All layers are normalized [2] and interconnected with residual connection [13] to ensure better gradient flow during training.

The decoder uses also attention to the encoder after each self-attentive layer. The decoder is autoregressive. In every time step, a new word is generated and the stack of all the layers applied on the text generated so far, including the newly generated word.

We use hyper-parameters and training strategy proposed by Popel and Bojar [23] who train the model in Tensor2Tensor. A vocabulary of 32,000 subwords is shared by the English encoder and Czech decoder. The network uses 16 self-attentive heads and a hidden layer of dimension 1,024. It is trained using the Adam optimizer [17] with the beta parameter set to 0.998 and the learning rate to 0.2 with 16,000 warmup steps, using a batch size of 1,500 and checkpoint averaging.

3.2 Dataset

We use CzEng 1.7[5] [5] Czech-English parallel corpus in a filtered version [23] which contains 57M pairs of parallel sentence pairs.

The model is validated on WMT13 test set and evaluated on WMT17 [6] test set from the news domain.

3.3 Evaluation

We evaluate the model on the WMT17 test set [6]. It is a test set that was used for system comparison in an annual competition in MT. Unlike the other 2 tasks which are rarely solved for Czech, English-to-Czech translation is annually evaluated within the WMT competition where it serves as an example of a highly inflected language.

The quantitative results are in Table 3, examples of the outputs in Table 2. As far as we know, this is the best publicly reported MT system for English-to-Czech translation.

Our best performing model was obtained by training for 8 days on 8 GPUs.

4 Image Captioning

In image captioning, the task is to provide a short textual description of a given image – i.e., the input for the task is an image (a two-dimensional matrix of bits, where each bit is represented by the values of its red, green and blue channel), and the output is a caption (a sequence of words).

As the Czech image captioning dataset is very new, we believe to be the first ones to train models for the image captioning task for Czech.

4.1 Architecture

We re-implement an attentive architecture by Xu et al. [29]. The model uses pre-trained convolutional map from networks for image classification on the ImageNet dataset [8], these are used as input to a RNN decoder with attention mechanism [3] originally introduced in context of MT.

Image features are extracted with Resnet50 v2 ($8 \times 8 \times 2048$) [13], captions are tokenized and truecased Moses style [18]. We use an RNN decoder [3] with conditional GRU [10] with dimensionality 1024, and our word embeddings have 500 dimensions. For Czech experiments, we use a vocabulary of 5,521 tokens, i.e., tokens that appear at least four times in the training data. For English, we use a vocabulary of 7,752 tokens appearing at least 5 times.

The model is optimized using the Adam optimizer [17] with default parameters and mini-batch size 64. Because we cannot rely on an extrinsic evaluation metric, we perform early stopping on reference captions perplexity.

At the inference time, we use a beam search of width 5 with length penalty 1.0 [28].

4.2 Dataset

We use a recently acquired Czech version of the Multi30k dataset [9] which contains translations of the originally English captions from the Flickr30k dataset [22].

The dataset uses 29,000 images for training, 1,014 for validation and 1,000 for testing. Unlike the original Flickr30k dataset which contains 5 independent descriptions for each image, we only have one Czech sentence for each image.

This means we can have only have one reference sentence for the evaluation which makes the evaluation less robust than in case of English.

4.3 Evaluation

Image captioning is usually evaluated using metrics originally developed for machine translation.

There is only one reference in the dataset, while the standard is to evaluate with BLEU [21] or METEOR [4] score against 6 references. In MT, 4 references are the standard [21], and 1 reference is typical in practice. In image captioning, the captions are quite short, and there is a much higher degree of freedom, which is why as many as 6 references are typically used. With only 1 reference available, BLEU cannot be reliably used here. However,

[5] http://ufal.mff.cuni.cz/czeng/czeng17

source:	The next chance won't come until winter.
system output:	Další příležitost přijde až v zimě.
reference:	Další šance přijde až v zimě.
source:	All private correspondence and images should remain private.
system output:	Veškerá soukromá korespondence a obrazy by měly zůstat soukromé.
reference:	Všechna soukromá korespondence a všechny soukromé obrázky by soukromé měly zůstat.

Table 2: An example of the outputs of the MT system.

cs output:	Skupina lidí stojí ve sněhu.
cs reference:	Skupina lidí stojící před iglú.
en output:	A group of people are standing in front of a building.
en reference:	A group of people wearing snowshoes, and dressed for winter hiking, is standing in front of a building that looks like it's made of blocks of ice. The people are quietly listening while the story of the ice cabin was explained to them. A group of people standing in front of an igloo. Several students waiting outside an igloo.

Figure 1: An example of an output of the image captioning system.

model	BLEU
Popel and Bojar [23] (ours)	23.8
WMT17 winner [25]	22.8
Google Translate [28]	20.8

Table 3: Qualitative evaluation of the MT model.

model	BLEU	METEOR	chrF3
Xu et al. [29]	19.1	18.5	—
ours (English)	19.7	17.0	0.17
ours (Czech)	2.3	7.2	0.14

Table 4: Quantitative results of the image captioning models.

as more references are not available, we use the chrF3 metric [24], which is based on character *n*-grams rather than word *n*-grams, and has thus a higher chance of providing at least somewhat useful evaluation scores (even though we note that they are still very unreliable).

We believe our work to be the first to perform image captioning in Czech language. As can be seen in Table 4, the standard evaluation shows rather low scores for Czech. However, when investigating the data, we found the produced image labels to be usually correct, even if rather simple and generic. See Figure 1 for an example of an input image together with its captions produced by our system.

5 Conclusion

We implemented and trained models for English-to-Czech machine translation, sentiment analysis of Czech texts, and image captioning in Czech within Neural Monkey, using approaches reported to be state-of-the-art for other languages (typically English). We gathered and standardized existing datasets, adapted the Neural Monkey toolkit where necessary, and trained and tuned the tools. Our evaluation shows that the resulting tools reach or surpass state-of-the-art for all three tasks. Both the source codes and the trained models are available online under free licences.[6][7] The tools are also available as an online demo.[8]

As a future work, we plan to add more tasks, especially text summarization.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang

[6] https://github.com/ufal/neuralmonkey
[7] http://hdl.handle.net/11234/1-2839
[8] https://ufal.mff.cuni.cz/grants/lsd

Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[5] Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovickỳ, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. Czeng 1.6: enlarged Czech-English parallel corpus with processing tools dockered. In *International Conference on Text, Speech, and Dialogue*, pages 231–238. Springer, 2016.

[6] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[7] Tomáš Brychcín and Ivan Habernal. Unsupervised improving of sentiment analysis using global target context. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 122–128, Hissar, Bulgaria, September 2013. INCOMA Ltd. Shoumen, BULGARIA.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255, Miami, FL, USA, jun 2009. IEEE, IEEE Computer Society.

[9] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics, 2016.

[10] Orhan Firat and Kyunghyun Cho. Conditional gated recurrent unit with attention mechanism. https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf, May 2016. Published online, version `adbaeea`.

[11] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[12] Ivan Habernal, Tomáš Ptáček, and Josef Steinberger. 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*, pages 65–74, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, TODO, TODO 2016. IEEE Computer Society.

[14] Jindřich Helcl and Jindřich Libovický. Neural Monkey: An open-source tool for sequence learning. *The Prague Bulletin of Mathematical Linguistics*, (107):5–17, 2017.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[18] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007.

[19] Ladislav Lenc and Tomáš Hercig. Neural networks for sentiment analysis in Czech. In *Proceedings of the 16th ITAT: Slovenskočeský NLP workshop (SloNLP 2016)*, pages 48–55, Bratislava, Slovakia.

[20] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[22] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2641–2649. IEEE, 2015.

[23] Martin Popel and Ondřej Bojar. Training tips for the Transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110:43–70, April 2018.

[24] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[25] Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Ger-

mann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[26] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416, 2018.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 6000–6010, Long Beach, CA, USA, December 2017. Curran Associates, Inc.

[28] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[29] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2048–2057. JMLR.org, 2015.