Transfer learning for Czech Historical Named Entity Recognition

Helena Hubková[†‡]
[†] Department of Computer
Science and Engineering,
University of West Bohemia,
Plzeň, Czech Republic
hhubkova@kiv.zcu.cz

Pavel Král[†‡]
[‡] NTIS – New Technologies
for the Information Society,
University of West Bohemia,
Plzeň, Czech Republic
pkral@kiv.zcu.cz

Abstract

Nowadays, named entity recognition (NER) achieved excellent results on the standard corpora. However, big issues are emerging with a need for an application in a specific domain, because it requires a suitable annotated corpus with adapted NE tag-set. This is particularly evident in the historical document processing field.

The main goal of this paper consists of proposing and evaluation of several transfer learning methods to increase the score of the Czech historical NER. We study several information sources, and we use two neural nets for NE modeling and recognition.

We employ two corpora for evaluation of our transfer learning methods, namely Czech named entity corpus and Czech historical named entity corpus. We show that BERT representation with fine-tuning and only the simple classifier trained on the union of corpora achieves excellent results.

1 INTRODUCTION

Recently, named entity recognition (NER) achieved outstanding results on standard NER corpora. Particularly on English CONLL-2003 corpus 90% F-measure has been overcome, which is sufficient for several real applications.

However, big issues are emerging with a need for an application in the specific domain which requires an appropriately annotated corpus including adapted NE tag-set. This issue is particularly evident in the case of historical documents in Czech language on which we focus on the project "Modern Access to Historical Sources"[1]. Manual annotation of this corpus is very expensive and time consuming. Moreover, the presence of a linguist is necessary. We will use the information about named entities as additional metadata for information retrieval and knowledge extraction.

Transfer learning targets at reusing information obtained from one corpus to improve the results of a model learned on an analogous task with few resources. To overcome this issue, we propose and evaluate several transfer learning approaches to improve the results of the Czech historical NER when the annotated resources are limited. The following information sources are considered and studied for this task:

- pre-trained fastText[2] word vectors;
- pre-trained word2vec[3] word vectors;
- pre-trained Slavic BERT[4] contextual text representation;
- Czech contemporary NE corpus from different domain.

We employ two neural based models, namely recurrent Bidirectional long short-term memory (BiLSTM) (Graves et al., 2005) and Bidirectional encoder representations from transformers (BERT) model (Devlin et al., 2019) with a simple perceptron for NER modelling and recognition. We use two corpora for evaluation of our experiments.

Note that, to the best of our knowledge, this is the first attempt to employ transfer learning in the field of Czech historical NER.

2 RELATED WORK

Rodriguez et al. (Rodriguez et al., 2018) presented reproduction paper focused on transfer learning for entity recognition. They compared seven new corpus pairs results and other researches that were

[1] http://www.portafontium.eu/

[2] https://fasttext.cc/

[3] https://code.google.com/p/word2vec/

[4] https://github.com/deepmipt/Slavic-BERT-NER

published previously. They showed that if there is a small labelled target dataset, the simpler approaches work better in compare to neural transfer approaches that work better for larger labelled target data set. They reached an F1 score of 71.81% for the period 1814–1817 and 70.35% for the period 1685–1691.

In term of NER in historical texts, different methods have been applied for English so far; rule-based NER (Grover et al., 2008), Maximum entropy Markov model and Conditional random fields (Packer et al., 2010). Different tools for NER (Rodriguez et al., 2012) as OpenNLP, Stanford NER, AlchemyAPI and OpenCalais are available. NER for historical newspapers was researched by Mac Kim and Cassidy (2015) (English), Neudecker Neudecker (2016) (Dutch, French, German) and Kettunen et al. (2016) (Finnish). In case of historical newspapers, Stanford NER (Finkel et al., 2005) was applied to the 155 million OCRed articles from historical Australian newspapers by Sunghwan and Cassidy (Mac Kim and Cassidy, 2015) and they described how the data can be exploited using a clustering method. Moreover, Neudecker (Neudecker, 2016) created an open corpus for NER in Dutch, French and German based on OCRed historical newspapers as part of the Europeana Newspapers project,[5] using Stanford NER for German. Similarly, (Kettunen et al., 2016) evaluated NER tools for Finnish using OCRed Finnish historical newspaper collection Digi. Transfer learning for NER was implemented by Lee et al. Lee et al. (2018), similarly, for historical German NER by Riedl and Padó (2018) and Schweter and Baiter (2019).

Transfer learning for NER was implemented by Lee et al. (Lee et al., 2018) using artificial neural nets for two different datasets of patient note de-identification. They demonstrated that an ANN model trained on large labeled data set could be transferred to get state-of-the-art results on the datasets with small number of labels. Transfer learning for historical NER was investigated by Riedl and Padó (Riedl and Padó, 2018). They compared different NER models and methods for both contemporary German (large datasets) and Historic German (small datasets). They concluded that the best performance has BiLSTM model with a CRF as a top layer if enough data is available. On the other hand, the BiLSTM model using transfer learning showed that it is more effective for small data. They trained the model with large datasets of contemporary German and then they tunned on small historical ones. More recently, Schweter and Baiter (Schweter and Baiter, 2019) applied the *contextual string embeddings* (Akbik et al., 2018) (Flair) for German Historic NER. They also used synthetic masked language modelling (SMLM) that randomly adds noise during the training in comparison to the masked language modelling in BERT by Devlin et al. (Devlin et al., 2019). They showed that pre-trained models on specific datasets can reach state-of-the-art results in the case of Historic German. However, the SMLM approach showed the second best results.They also experimented with pre-trained fastText embeddings.

Recently, NER for contemporary Czech was researched by Straka et al. (2019) using BERT (Devlin et al., 2019) and Flair (Akbik et al., 2018). Similarly, Arkhipov et al. (2019) resented multilingual NER for Russian, Bulgarian, Czech and Polish.

In term of NER for contemporary Czech, Straka et al. (Straka et al., 2019) recently presented their sequence-to-sequence model to evaluate BERT (Devlin et al., 2019) and Flair (Akbik et al., 2018) and their combination on Czech named entity corpus (CNEC) versions 1.1 and 2.0[6]. For CNEC 1.1, they reached 87.62% F1-score using Flair, 89.85% using BERT and 89.91% using both of them. For types of CNEC 2.0., they achieved 81.65% F1-score for Flair, 86.23% for BERT and 85.52% for both.

Moreover, Arkhipov et al. (Arkhipov et al., 2019) presented multilingual named entity recognition in Russian, Bulgarian, Czech and Polish (Asia Bibi datasets from BSNLP 2019 Shared Task) using BERT model and additional word-level CRF layer. This approach reached state-of-the-art results: 93.9 F1 score for Czech, 87.3 for Russian, 87.2 for Bulgarian and 93.2 for Polish, respectively.

In the case of text embeddings, Akbik et al. proposed a pre-trained model of *contextual string embeddings* (Flair) for NER that considers words as sequences of characters. They experimented with a BiLSTM-CRF model proposed by Huang et al. (Huang et al., 2015) and different approaches to word embeddings. They extended the model by adding a concatenation of pre-trained static word embeddings with contextual ones and a concate-

[5] http://www.europeana-newspapers.eu/

[6] https://ufal.mff.cuni.cz/cnec

nation of task-trained character features with contextual string embeddings. They reached 93.09% F1-score for English and 88.32% for German with this model configuration for CoNLL2003 shared task.

3 CORPORA

We experimented with two corpora: Czech named entity corpus (CNEC) and Czech historical named entity corpus (CHNEC). CNEC corpus contains almost 9,000 sentences and more than 35,000 occurrences of the Czech named entities. The corpus uses two-level NEs annotation scheme and the first-level contains 10 main NE types and the second-level is composed of 62 NE subtypes. To be able to map NEs from CNEC to CHNEC, we use only five NE types from the first annotation level which are same for both corpora.

CHNEC[7] contains 73,647 tokens and 4,017 named entity occurrences. The corpus was created from Czech historical newspaper *Posel od Čerchova* from second half of 19th century and distinguishes five NE types: *Personal names*, *Institutions*, *Geographical names*, *Time expressions* and *Artifact names/Objects*. The corpus is encoded in IOB format (Ramshaw and Marcus, 1995), where *B* represents one-token entity or the beginning of multi-token named-entity, *I* inside tokens of multi-token named-entity and *O* stands for all tokens that are not a named-entities.

4 METHODS

4.1 Models and Representations

4.1.1 BiLSTM with Word-level Embeddings

The first approach uses BiLSTM model with word-level representation of the sentences. We used similar network structure and similar hyperparameters as presented by Hubková et al. (2020) (Table 1) with two different word representation methods: fastText and word2vec.

Note that we use BiLSTM model with randomly initialized word embeddings as a baseline. This approach does not consider any transfer because the embeddings are learned during the training of the network only of the available training data.

4.1.2 BERT with Perceptron

The second method uses BERT model (Devlin et al., 2019) for representation of the text and a simple single-layer perceptron (SLP) with only one softmax layer is used for NE recognition. The main advantage of this approach in comparison with the previous one is that BERT considers the different word meaning when used in the different context.

Hyper-parameter	Range	Final
LSTM state #	[100; 500]	250
LSTM layer #	[1; 3]	1
Learning rate	[0.001; 0.01]	0.004
Epochs	[60;120]	80
Dropout	[0.25; 0.85]	0.65

Table 1: Overview of hyper-parameter optimization

This model uses unlabeled data to pretrain deep bidirectional representations by jointly conditioning on both left and right context in all layers. Sequences of word tokens (or subtokens) in the sentence are used as an input and the outputs are class probabilities among the classes.

4.2 Transfer Learning

4.2.1 Transfer from Embedding Word Vectors

The first transfer learning approach is focused on the embedding vectors. Our embedding vectors are built using different models, and thus they kept different information. FastText considers word and also sub-word units, therefore it should encode semantic and syntactic information as well. However, word2vec is trained properly using word tokens, hence it includes mostly semantic information.

The initial embeddings are learned on huge unlabelled text corpora coming from different domains and containing different language (contemporary Czech instead of historical one). For that reason, we assume that the further fine-tuning of these embeddings on the target data should improve the final NE recognition score. Therefore, in this approach, we compare fastText and word2vec and embeddings with two scenarios. The first one uses only static embeddings without a subsequent training and the second one tries to adjust these vectors into our task during network training.

4.2.2 Transfer from BERT Representation

The second approach is based on the transfer from BERT representation. In order to have as much precious representation as possible, we use pre-trained Slavic BERT proposed by Arkhipov et al. (2019). This model is based on a multilingual BERT model and fine-tuned with Czech, Russian, Bulgarian and

[7] http://chnec.kiv.zcu.cz/

Polish data. We assume that the further fine-tuning of this representation on the target data will still improve the final NE score. Therefore, in this approach, we perform another fine-tuning by training on our historical data.

4.2.3 Transfer from Different Corpus

The third approach assumes that different NER corpora (see Section 3) should include complementary information and the usage of these together will improve the final score of the target task. The following scenarios are considered, evaluated and compared:

- training only on CHNEC corpus (to show the impact of the approaches below);
- training only on CNEC corpus (to show the results when the target annotated data are not available);
- union of both corpora and training on this large dataset;
- initial training on the CNEC corpus and fine-tuning on the target one (CHNEC).

4.3 Cross-corpus Method

As a cross-corpus method we mean that we trained BiLSTM or BERT model with CNEC training data set (source corpus) and models and we tested on CHNEC test data (target corpus). This approach showed if bigger corpora for contemporary language itself can be used for tagging a smaller historical texts.

We experimented with token classification PyTorch module for NER by Wolf et al. (2019) and we used pre-trained Slavic BERT model.

BERT token classification method has a linear layer on top of the hidden-states output and this model is available through PyTorch[8] module. Pre-trained Slavic BERT is based on BERT model by Devlin et al. (2019) and extended by word-level CRF layer. The model is tunned on four Slavic languages Russian, Bulgarian, Czech and Polish.

5 Evaluation and Results

Table 2 and Table 3 show the results of our experiments. We use the standard precision, recall and macro-averaged F1-score (Powers, 2011) metrics for evaluation. In all cases, we calculate the final score on the testing part of CHNEC corpus. Qualitative analysis in Section 5.1 is based on the observed linguistics phenomena in both development and test data sets.

The first part of Table 2 presents the results of our approaches dealing with the transfer of embedded word vectors. BiLSTM model trained only on CHNEC corpus is used for NE recognition. The results show that fastText representation brings significantly better results than word2vec one. These results further illustrate that the fine-tuning of the embeddings has only a positive impact in the word2vec case, and unfortunately, it does not bring any improvement in the case of the fastText representation. This behaviour should be justified that fastText word representation corresponds better to our task, and our training data are too small and differ from the testing set for the further improvement of the model. Based on these results, we will use for the following experiments only fastText fixed embeddings.

The second part of the table shows the experiments using BiLSTM model with different approaches for training. These results show that it is possible to obtain F-measure 45% using only the different corpus (any additional annotation is not required). Moreover, it has been demonstrated that the transfer from CNEC into CHNEC does not bring any positive impact on the final NER.

The last part of Table 2 shows the results of our transfer learning approaches using BERT representation with the simple single-layer perceptron as a classifier. The impact of the different training approaches is also considered. These results show clearly that BERT representation with fine-tuning and only a simple classifier brings significantly better results that all the approaches evaluated previously. This should be explained by the fact that word context representation is much more accurate than the word-level one. This experiment further illustrates that the additional data coming from another NER corpus is beneficial to improve the final NER score by 2%.

Another observation is that the values of the precision and recall are balanced, however in the previous series of experiments, these values differ significantly.

The previous experiment does not consider the individual NE types. However, this information could be very interesting for further improvement of the training strategy and the model itself. We assume, that the size of the training data and the

[8] https://pytorch.org/

No.	Approach	Training data	Precision	Recall	F1
1	BiLSTM (baseline)	CHNEC	0.63	0.52	0.57
	BiLSTM + fastText fixed		0.76	0.70	0.73
	BiLSTM + fastText fine-tuned		0.65	0.60	0.63
	BiLSTM + word2vec fixed		0.67	0.45	0.54
	BiLSTM + word2vec fine-tuned		0.61	0.55	0.58
2	BiLSTM + fastText fixed	CNEC	0.53	0.39	0.45
	BiLSTM + fastText fixed	CNEC + CHNEC	0.63	0.71	0.67
	BiLSTM + fastText fixed	CNEC → CHNEC	0.66	0.71	0.69
3	SLP + Slavic BERT fine-tuned	CHNEC	0.79	0.81	0.80
	SLP + Slavic BERT fine-tuned	CNEC	0.61	0.53	0.57
	SLP + Slavic BERT fine-tuned	CNEC + CHNEC	**0.81**	**0.84**	**0.82**

Table 2: NER results of the different approaches evaluated on the testing set of CHNEC corpus (best results in bold)

NE Class	Precision	Recall	F1-score	NE # in Total	NE # in Test
Geographical names	0.91	0.93	0.92	1104	137
Artifact names/Objects	0.85	0.86	0.86	829	101
Time expressions	0.79	0.75	0.77	506	55
Personal names	0.72	0.76	0.74	1292	91
Institutions	0.57	0.63	0.59	286	37

Table 3: Results of the individual NE types using the best model and training scenario: *SLP + Slavic BERT fine-tuned on CNEC + CHNEC*. The last two columns show the NE numbers in the whole and in the testing part of the CHNEC corpus.

complexity of the target NE will influence the results of this particular class.

Therefore, Table 3 shows the recognition results of the individual NE types. We also report the appropriate NE numbers in the whole corpus and in the testing part. We use the best transfer learning strategy and model identified previously, i.e. *SLP + Slavic BERT fine-tuned trained on CNEC + CHNEC corpora*. The results of this table confirm our assumption reported above. This table shows that the best recognition score, F1-score more than 90%, is obtained for *Geographical names*, which has high amount of the training occurrences and relatively simple structure. On the other hand, the lowest F1-score 0.59% has been obtained for *Institutions*, which has the lowest number of training data and very complex structure (irregular multi-token entities). This experiment thus confirms our assumption.

5.1 Qualitative Analysis

If the different models produce different errors, the combination of the models should bring the improvement of the final score. Therefore, we analyze deeper the type of errors of the different models and the learning approaches on the randomly selected sample containing about 100 randomly selected sentences per model and approach.

If we train BiLSTM model only with CNEC data, the model showed that this method correctly tagged especially *geographic names* and *personal names*, e.g.: names of towns such as *Prahy* or *Horšovský Týn*; names of persons such as *Vojtěcha Bittnara* or *Petr Bedřich Florian*. The *time expressions* were tagged correctly in case that the format of the time in CNEC was similar to format in CHNEC, e.g. *20 . února 1775* ("20th February 1775"). Similarly, the named entity *Sokol* (name of a sport institution) was tagged correctly as this named entity occurs in both CNEC and CHNEC. However, the other NE types were rather erroneous.

On the other hand, if we compare these results with a basic BiLSTM model that was trained only on in-domain historical data, we can see that the specific language expressions that occur in historical CHNEC texts are not tagged. CHNEC corpus contains a number of abbreviations, e.g. *26 . června t . r .* ("26th June this year") or *c . k . okresní finanční řediteství v Plzni* ("Imperial-Royal District Financial Directorate in Pilsen"). Moreover, using dots in the named entities in this corpus is inconsistent, it means that it contains both *29 . June* and *29 , June* or name of person *Jos . Kralovec*.

Then, we analyze the results of the approaches which use BERT representation. Generally, BERT improves the results with its ability to correctly tag NEs that do not occur in the training data. As we fine-tuned the Slavic BERT model using the combination of CNEC and CHNEC, the model overcomes the problems with abbreviations mentioned above. However, some false positive cases occur as well, e.g. *c . k . okresní hejtman* ("Imperial-Royal district governor") was tagged as *Institution*.

Geographical names that occur more frequently in CHNEC than most other NE types reached 0.92 F1-score in comparison to *Institutions* that occur only 286 times in the whole CHNEC and reached 0.59 F1-score.

Next to the pure occurrences in the corpora, the names of the *Institutions* differ between historical and contemporary language a lot as many institutions do not exist in the present anymore, e.g. *spořitelny kr . města Domažlic* ("savings bank of the royal town of Domažlice"). From this point of view, *Time expressions* or *Personal names* are more stable in time. The *Institution* are also usually long multi-word expressions (e.g. *" všeobecné úvěrní a obchodní banky "* ("general credit and commercial bank") in contrast to shorter and more consistent *Time expressions* and *Personal names*. This fact also complement the previous justifications of the lowest recognition score for the class *Institutions*.

6 Conclusions

In this paper, we proposed and evaluated several transfer learning approaches in order to improve the results of the Czech historical NER. We considered and studied the following information sources: pre-trained fastText and word2vec word representations, pre-trained Slavic BERT contextual text representation and another NE corpus from different domain. We used two popular models, namely recurrent BiLSTM and BERT with a simple perceptron for NER modelling and recognition. We have shown that fastText representation gives significantly better results than word2vec model. It has been also demonstrated that the transfer from CNEC into CHNEC with BiLSTM model does not improve the final NER score. We have further presented that BERT representation with fine-tuning and only the simple classifier trained on the union of corpora brings the best results (F-measure 82%). Based on the analysis of the errors, we can also conclude that the combination of the different models / approaches would not bring any further improvement.

Acknowledgements

This work has been partly supported from ERDF "Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)" (no.: CZ.02.1.01/0.0/0.0/17_048/0007267) and by Grant No. SGS-2019-018 Processing of heterogeneous data and its specialized applications.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*, pages 799–804. Springer.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. In *LREC 2008*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Helena Hubková, Pavel Král, and Eva Pettersson. 2020. Czech historical named entity corpus v 1.0. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4460–4467, Marseille, France. European Language Resources Association.

Kimmo Kettunen, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala, and Laura Löfberg. 2016. Old content and modern tools - searching named entities in a finnish ocred historical newspaper collection 1771-1910. *CoRR*, abs/1611.02839.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sunghwan Mac Kim and Steve Cassidy. 2015. Finding names in trove: named entity recognition for australian historical newspapers. In *Australasian Language Technology Association Workshop 2015*, volume 13, pages 57–65. Australasian Language Technology Association.

Clemens Neudecker. 2016. An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Thomas L. Packer, Joshua F. Lutes, Aaron P. Stewart, David W. Embley, Eric K. Ringger, Kevin D. Seppi, and Lee S. Jensen. 2010. Extracting person names from diverse and noisy OCR text. In *AND*, pages 19–26. ACM.

DMW Powers. 2011. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Martin Riedl and Sebastian Padó. 2018. A named entity recognition shootout for German. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125, Melbourne, Australia. Association for Computational Linguistics.

Juan Diego Rodriguez, Adam Caldwell, and Alexander Liu. 2018. Transfer learning for entity recognition of novel classes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1974–1985, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kepa J. Rodriguez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. Comparison of named entity recognition tools for raw ocr text.

Stefan Schweter and Johannes Baiter. 2019. Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.

Milan Straka, Jana Strakova, and Jan Hajič. 2019. 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*, pages 137–150, Cham / Heidelberg / New York / Dordrecht / London. Springer International Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.