Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7386–7391
Marseille, 20-25 June 2022

Pre-training and Evaluating Transformer-based Language Models for Icelandic

Jón Friðrik Daðason, Hrafn Loftsson
Department of Computer Science
Reykjavik University, Iceland
{jond19, hrafn}@ru.is

Abstract

In this paper, we evaluate several Transformer-based language models for Icelandic on four downstream tasks: Part-of-Speech tagging, Named Entity Recognition. Dependency Parsing, and Automatic Text Summarization. We pre-train four types of monolingual ELECTRA and ConvBERT models and compare our results to a previously trained monolingual RoBERTa model and the multilingual mBERT model. We find that the Transformer models obtain better results, often by a large margin, compared to previous state-of-the-art models. Furthermore, our results indicate that pre-training larger language models results in a significant reduction in error rates in comparison to smaller models. Finally, our results show that the monolingual models for Icelandic outperform a comparably sized multilingual model.

Keywords: Language Models, Transformer, Evaluation, Icelandic

1. Introduction

Pre-trained language models have obtained state-of-the-art performance on a wide variety of Natural Language Processing (NLP) tasks, including Question Answering (QA), Named Entity Recognition (NER), Part-of-Speech (POS) tagging and Automatic Text Summarization (ATS) (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019). Such models are first *pre-trained* on a large, unannotated corpus on an unsupervised task, such as recovering tokens which have been randomly masked. Once pre-trained, the model can be *fine-tuned* on smaller, annotated datasets for more practical tasks – also known as *downstream tasks* in this context.

It is well established that increasing the size of the pre-training corpus has a positive impact on downstream performance (Liu et al., 2019; Lan et al., 2019; Raffel et al., 2020). As a result, the amount of pre-training data used by recent models has grown exponentially, from 800 million tokens in 2018 (GPT) (Radford et al., 2018) to over one trillion tokens in 2020 (T5) (Raffel et al., 2020). This is several orders of magnitude more data than is available for low and medium-resource languages.

Another option is to use multilingual models, for which pre-training data is plentiful. Multilingual Transformer models have, in some instances, achieved comparable or even better results than monolingual models (Conneau et al., 2020; Xue et al., 2021). However, this is not always the case, especially for low-resource languages, where monolingual models tend to obtain better results (Pyysalo et al., 2020). One reason may be that low-resource languages typically make up an insignificant portion of the pre-training corpus, which can consist of text from over 100 languages.

In this paper, we present our first results with pre-training and evaluating monolingual Transformer-based models for Icelandic. We pre-train four types of ELECTRA (Clark et al., 2020) and ConvBERT (Jiang et al., 2020) models and evaluate them on four downstream tasks: POS tagging, NER, Dependency Parsing (DP), and ATS. We compare our results to previous state-of-the-art models and baselines. On all our downstream tasks, we find that the Transformer-based models obtain better results, often by a large margin, compared to previously published state-of-the-art models. Furthermore, our results show that the larger models obtain significantly better results than the smaller models. Finally, we find that monolingual models for Icelandic outperform a comparably sized multilingual model.

The main contributions of our work are the following:

- Pre-training of four types of ELECTRA and ConvBERT models for Icelandic. These types of Transformer models have not been pre-trained and published before for Icelandic[1].
- Fine-tuning and evaluation of the above mentioned models, in addition to an already existing Icelandic RoBERTa model(Liu et al., 2019; Snæbjarnarson et al., 2022) and the multilingual mBERT model(Devlin, 2018), on four different downstream tasks.

The rest of the paper is structured as follows. We discuss related work in Section 2 and the pre-training of our models in Section 3. In Section 4, we present the four downstream tasks, and the evaluation results in Section 5. Finally, we conclude in Section 6.

[1] The resulting models are publicly available on the Hugging Face model repository – see Section 3. For training the models, we followed the guidelines available at `https://github.com/stefan-it/turkish-bert`.

2. Related Work

BERT (Devlin et al., 2019) is a bidirectional Transformer-based model, pre-trained with the *masked language modelling* (MLM) task, where a certain percentage of tokens are replaced with a special mask token, and the model attempts to predict its original form. Unlike a next token prediction task used by generative language models, this allows the model to use contextual information from both directions, making it more suitable for classification tasks. BERT obtained state-of-the-art results on all 11 downstream tasks it was evaluated on.

Many optimizations to the Transformer architecture have been proposed since its introduction. The RoBERTa model obtains significantly better downstream performance than BERT by increasing the amount of pre-training data, the number of pre-training steps and the batch size (Liu et al., 2019).

Wu and Dredze (2020) suggest that the MLM pre-training task is data-inefficient, as the model only makes predictions for the relatively small number of tokens which are masked. In the case of BERT, only 15% of the tokens in the pre-training corpus are masked by default. Clark et al. (2020) propose a more efficient pre-training task known as *replaced token detection* (RTD), where a certain percentage of tokens are replaced with other tokens from the pre-training corpus, avoiding the use of a pre-training specific mask token. The model, called ELECTRA, then makes a binary prediction for each token, attempting to determine whether it is original or if it was replaced. The model also includes several other optimizations, such as weight sharing and dynamic masking, where tokens are masked or replaced dynamically during training rather than during a pre-processing step.

ConvBERT (Jiang et al., 2020) is an implementation of the ELECTRA model where a span-based dynamic convolution operator has been integrated into the self-attention mechanism. This improvement results in slightly better downstream performance while significantly increasing the computational efficiency of the model at larger sizes.

The authors of BERT later released mBERT[2], a BERT model which was trained on a corpus consisting of all articles from the 104 largest languages on Wikipedia. While mBERT performs reasonably well for most languages, Pyysalo et al. (2020) show that in a majority of cases, a monolingual BERT model, trained on Wikipedia articles, obtains similar or better downstream performance. Additionally, Wu and Dredze (2020) show that on the 30% of its smallest languages in the pre-training corpus, mBERT is outperformed by bidirectional long short-term memory (BiLSTM) models.

Many language-specific models have been released in recent years, including IceBERT for Icelandic (Snæbjarnarson et al., 2022), FinBert for Finnish (Virtanen et al., 2019), CamemBERT for French (Martin et al., 2020) and RobBERT for Dutch (Delobelle et al., 2020). In all cases, these models outperform both previous state-of-the-art methods as well as mBERT.

3. Pre-training

We pre-trained four models:

- An ELECTRA-Small model[3] consisting of 14M parameters and requiring 8 hours to train on a TPU v3-8 node.
- An ELECTRA-Base model[4] consisting of 110M parameters and requiring 10 days to train on a TPU v3-8 node.
- A ConvBERT-Small model[5] consisting of 14M parameters and requiring 15 hours to train on a TPU v3-8 node.
- A ConvBERT-Base model[6] consisting of 106M parameters and requiring 13 days to train on a TPU v3-8 node.

All four models were pre-trained on the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018) using default settings. The IGC is an unannotated corpus containing 1.69B tokens from various domains, including news articles, parliamentary speeches, books and blogs. We used a byte-pair encoding (BPE) tokenizer with a vocabulary size of 32k, maintaining the original casing and accents.

4. Fine-tuning

We fine-tuned and evaluated our models on four downstream tasks: POS tagging, NER, DP, and ATS. The datasets, experimental settings and evaluation metrics are described in the following sections.

4.1. Part-of-speech tagging

For POS tagging, the models were fine-tuned and evaluated on the MIM-GOLD corpus (Loftsson et al., 2010), which consists of approximately one million tokens that have been semi-automatically annotated with POS tags. The fine-tuning was performed using the Transformers library (Wolf et al., 2020), with a batch size of 16 and a learning rate of 5e-5 for 20 epochs.

We report tagging accuracy using the standard 10-fold cross-validation sets for MIM-GOLD (Barkarson et al., 2021).

[2] `https://github.com/google-research/bert/blob/master/multilingual.md`

[3] `https://huggingface.co/jonfd/electra-small-igc-is`

[4] `https://huggingface.co/jonfd/electra-base-igc-is`

[5] `https://huggingface.co/jonfd/convbert-small-igc-is`

[6] `https://huggingface.co/jonfd/convbert-base-igc-is`

4.2. Named entity recognition

We fine-tuned and evaluated NER models on the MIM-GOLD-NER corpus (Ingólfsdóttir et al., 2020a), a version of MIM-GOLD which has been manually annotated with eight types of named entities. The models were trained using the Transformers library for 10 epochs with the same settings as the POS model. We report the overall entity-level F_1 scores using 10-fold cross-validation. Currently, there are no standard 10-fold cross-validation splits available for MIM-GOLD-NER. Therefore, we created a stratified 10-fold split, with each fold containing roughly the same proportion of NER tags.

4.3. Dependency parsing

For DP, the models were fine-tuned and evaluated on the Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2011), which consists of approximately one million tokens that have been manually annotated with constituents. We used a version of IcePaHC which has been converted to the Universal Dependencies format (Arnardóttir et al., 2020).

The training was performed using DiaParser, a BiLSTM-based biaffine dependency parser which extracts contextual word embeddings from a Transformer-based language model (Attardi et al., 2021). The models were fine-tuned using default settings. We report the labeled attachment score (LAS) for each model.

4.4. Automatic text summarization

Finally, we fine-tuned and evaluated the models on ATS using the IceSum corpus (Daðason et al., 2021), which contains 1,000 Icelandic news articles that have been manually annotated with extractive summaries. Each annotated summary consists of a list of full sentences or independent clauses from the corresponding news article. The models were fine-tuned using the TransformerSum[7] library for Python, which implements the BertSum extractive text summarization model (Liu and Lapata, 2019).

We report results using the ROUGE metric, which measures the ratio of overlapping n-grams between the target summary and the generated summary. We calculate the ROUGE score by comparing the first 100 words of the generated summary against the target summary. We report ROUGE-2 recall scores, averaged over five runs. Like Daðason et al. (2021), when creating the training data, we use an oracle to label each sentence in the original document, greedily maximizing the ROUGE-2 recall score until the summary exceeds 100 words in length. We used 70% of the corpus for training, with the rest being split equally between validation and test sets. The models were fine-tuned for 5 epochs using a linear classifier. Sentence embeddings were obtained by averaging token vectors within a sentence. Otherwise, default settings are used.

5. Results

The results for each task is detailed in the following sections. In addition to our pre-trained models, we include results obtained by IceBERT-IGC[8] (Snæbjarnarson et al., 2022), an Icelandic RoBERTa-Base model which was also pre-trained on the IGC. Furthermore, we include results obtained by the multilingual mBERT model.

5.1. Part-of-speech tagging

The tagging accuracy of the fine-tuned models is shown in Table 1. We compare our results against those obtained by ABLTagger, a BiLSTM-based model which is augmented with a morphological lexicon (Steingrímsson et al., 2019). Our findings show that the Transformer models outperform the BiLSTM-based approach by a significant margin, reducing the error rate by as much as 54%.

There is virtually no difference between the tagging accuracy of the ELECTRA and ConvBERT models. However, the results show that the larger monolingual models significantly outperform their smaller versions, with ConvBERT-Base reducing the error rate of ConvBERT-Small by approximately 28%. This demonstrates the benefits of training larger models even in settings where pre-training data is not abundant (i.e., not in the order of several billion tokens or more).

Model	Accuracy
ABLTagger	95.15%
mBERT	96.38%
ELECTRA-Small	96.84%
ConvBERT-Small	96.88%
IceBERT-IGC	97.37%
ELECTRA-Base	97.72%
ConvBERT-Base	97.75%

Table 1: Accuracy obtained on POS tagging.

5.2. Named entity recognition

The overall entity-level F_1 score of each model is shown in Table 2. We also include results for a BiLSTM-based model with pre-trained word embeddings, which had previously obtained state-of-the-art results on the MIM-GOLD-NER dataset (Ingólfsdóttir et al., 2020b). The model was originally trained for 100 epochs with early stopping, using 80% of the data for training, 10% for validation and 10% for testing, obtaining an F_1 score of 83.90%. Since we do not use a validation split, we instead report the average F_1 score obtained after 10 epochs on each fold. Our evaluation shows that the BiLSTM model obtains an F_1

[7] https://github.com/HHousen/TransformerSum

[8] https://huggingface.co/mideind/IceBERT-igc

score of 87.07%, which suggests that the original split may be considerably more challenging than ours. We find that the Transformer-based models outperform the previous state-of-the-art by a much wider margin than for POS tagging, reducing the error ratio by 55%. Our results also show that the multilingual mBERT model only outperforms the much smaller ELECTRA-Small model and the BiLSTM baseline.

As was the case for POS tagging, our results for NER demonstrate the importance of pre-training larger models. The F_1 score of ConvBERT-Base compared to ConvBERT-Small is equivalent to a 26% reduction in the error rate.

Model	F_1
BiLSTM	87.07%
ELECTRA-Small	91.23%
mBERT	91.31%
ConvBERT-Small	92.03%
IceBERT-IGC	93.04%
ELECTRA-Base	93.75%
ConvBERT-Base	94.14%

Table 2: F_1 scores obtained on NER.

5.3. Dependency parsing

The LAS for each model is listed in Table 3. The results show a negligible difference between the smaller models, while ELECTRA-Base obtains the highest score of the four larger models. The difference in LAS between ConvBERT-Base and ConvBERT-Small is equivalent to an 11% reduction in the error rate. The multilingual mBERT model obtains the lowest score.

Model	**LAS**
mBERT	82.94%
ConvBERT-Small	84.75%
ELECTRA-Small	84.90%
IceBERT-IGC	85.04%
ELECTRA-Base	86.20%
ConvBERT-Base	86.50%

Table 3: Label Attachment Score (LAS) obtained on DP.

5.4. Automatic text summarization

Table 4 shows the ROUGE-2 recall score of the models that were fine-tuned on the IceSum corpus. We include the score obtained by the Lede baseline, which creates a summary from the first few sentences of the news article. While extremely simple, this baseline obtains surprisingly good results when summarizing news articles. For Lede-100, we add sentences from the beginning of the document to the generated summary until it contains at least 100 words. We also show the score obtained by the oracle algorithm described in Section 4.4, which denotes the highest possible score that each model can obtain, when evaluated against the original annotated summaries. For comparison purposes, we limit the length of each generated summary to 100 words. Additionally, we include results obtained by a bidirectional, sequence-to-sequence model with attention proposed by Kedzie et al. (2018). The model is trained using the nnsum[9] library for 30 epochs. The model with the highest ROUGE-2 recall score on the validation set is saved. For the other models, we report the average score out of five runs.

The results show that the Seq2Seq model, which had previously obtained state-of-the-art results on extractive text summarization for Icelandic (Daðason et al., 2021), is outperformed by ELECTRA-Base and ConvBERT-Base. Only the Seq2Seq, ELECTRA-Base, and ConvBERT-Base models demonstrate a significant improvement over the Lede baseline. There is little to no difference between the quality of summaries generated by the other models.

Model	**ROUGE-2**
mBERT	69.09
Lede-100	69.14
IceBERT-IGC	69.14
ELECTRA-Small	69.29
ConvBERT-Small	69.36
Seq2Seq	70.42
ELECTRA-Base	71.04
ConvBERT-Base	71.09
Oracle	89.48

Table 4: ROUGE-2 recall scores obtained on ATS.

6. Conclusions

We have pre-trained and evaluated several different Transformer-based monolingual language models on four downstream tasks. We have shown that they outperform previous state-of-the-art models, as well as mBERT, a large multilingual Transformer model. Moreover, our results indicate that it is beneficial to put emphasis on pre-training larger monolingual models for Icelandic.

For future work, we intend to evaluate different tokenization algorithms and vocabulary sizes under low and medium-resource settings. Additionally, we plan to experiment with augmenting the IGC pre-training corpus with both monolingual data from online sources, as well as multilingual data from related languages. Finally, we will continue to experiment with different Transformer architectures and model sizes.

7. Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019-2023 (Nikulásdóttir et

[9] `https://github.com/kedz/nnsum`

al., 2020). The programme, which is managed and coordinated by Almannarómur[10], is funded by the Icelandic Ministry of Education, Science and Culture. This research was also supported with Cloud TPUs from Google's TPU Research Cloud (TRC).

8. Bibliographical References

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., and Steingrímsson, S. (2020). A Universal Dependencies Conversion Pipeline for a Penn-format Constituency Treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Attardi, G., Sartiano, D., and Simi, M. (2021). Biaffine Dependency and Semantic Graph Parsing for EnhancedUniversal Dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 184–188, Online. Association for Computational Linguistics.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

Daðason, J., Loftsson, H., Sigurðardóttir, S., and Björnsson, Þ. (2021). IceSum: An Icelandic Text Summarization Corpus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 9–14, Online. Association for Computational Linguistics.

Delobelle, P., Winters, T., and Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online, November. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Devlin, J. (2018). Multilingual BERT Readme Documentation.

Ingólfsdóttir, S. L., Guðjónsson, Á. A., and Loftsson, H. (2020b). Named Entity Recognition for Icelandic: Annotated Corpus and Models. In Luis Espinosa-Anke, et al., editors, *Proceedings of the 8th International Conference on Statistical Language and Speech Processing (SLSP 2020)*, pages 46–57, Cardiff, United Kingdom.

Jiang, Z.-H., Yu, W., Zhou, D., Chen, Y., Feng, J., and Yan, S. (2020). ConvBERT: Improving BERT with Span-based Dynamic Convolution. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12837–12848. Curran Associates, Inc.

Kedzie, C., McKeown, K., and Daumé III, H. (2018). Content Selection in Deep Learning Models of Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Liu, Y. and Lapata, M. (2019). Text Summarization with Pretrained Encoders. 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Loftsson, H., Yngvason, J. H., Helgadóttir, S., and Rögnvaldsson, E. (2010). Developing a PoS-tagged corpus using existing tools. In Francis M. Tyers Sarasola, Kepa et al., editors, *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, LREC 2010, Valetta, Malta.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.

Nikulásdóttir, A., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. (2020). Language technology programme for Icelandic 2019-2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France, May. European Language Resources Association.

[10] `https://almannaromur.is/`

Pyysalo, S., Kanerva, J., Virtanen, A., and Ginter, F. (2020). WikiBERT models: deep transfer learning for many languages. *arXiv preprint arXiv:2006.01538*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Ingólfsdóttir, S. L., Jónsson, H. P., Þorsteinsson, V., and Einarsson, H. (2022). A Warm Start and a Clean Crawled Corpus – A Recipe for Good Language Models. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Steingrímsson, S., Kárason, Ö., and Loftsson, H. (2019). Augmenting a BiLSTM Tagger with a Morphological Lexicon and a Lexical Category Identification Step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria, September.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wu, S. and Dredze, M. (2020). Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

9. Language Resource References

Barkarson, Starkaður and Andrésdóttir, Þórdís Dröfn and Hafsteinsdóttir, Hildur and Magnússon, Árni Davíð and Rúnarsson, Kristján and Steingrímsson, Steinþór and Jónsson, Haukur Páll and Loftsson, Hrafn and Sigurðsson, Einar Freyr and Rögnvaldsson, Eiríkur and Helgadóttir, Sigrún. (2021). *MIM-GOLD 21.05*. CLARIN-IS.

Daðason, Jón Friðrik and Loftsson, Hrafn and Sigurðardóttir, Salome Lilja and Björnsson, Þorsteinn. (2021). *IceSum - Icelandic Text Summarization Corpus*. CLARIN-IS.

Ingólfsdóttir, Svanhvít Lilja and Guðjónsson, Ásmundur Alma and Loftsson, Hrafn. (2020a). *MIM-GOLD-NER – named entity recognition corpus*. CLARIN-IS.

Wallenberg, Joel C. and Ingason, Anton Karl and Sigurðsson, Einar Freyr and Rúnarsson, Kristján and Rögnvaldsson, Eiríkur. (2011). *The Icelandic Parsed Historical Corpus (IcePaHC)*. CLARIN-IS.