

TRAINING A NEW GENERATION OF LTG LANGUAGE MODELS

David Samuel

Language Technology Group, University of Oslo

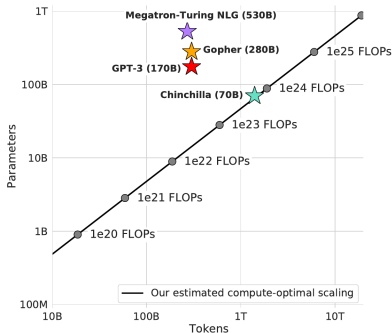




1. 🌱 **Pretraining on 100M-word corpus**
2. 🌲 New generation of Norwegian language models
3. 🤔 Surprise

Data bottleneck

- Language modeling is now in a place where simple scaling is not enough, we are bottlenecked by the amount of available training data.
- **It is time to experiment with new training approaches.**
- We propose to focus on improving language modeling strategies using small curated corpora.





Accessible and reproducible research

- Small, representative and openly available text corpus makes language models **easily reproducible and comparable**.



Accessible and reproducible research

- Small, representative and openly available text corpus makes language models **easily reproducible and comparable**.
- And without even sacrificing good performance.



Accessible and reproducible research

- Small, representative and openly available text corpus makes language models **easily reproducible and comparable**.
- And without even sacrificing good performance.
- We use the **British National Corpus** (Consortium, 2007) as such dataset



Brief description

- We use the BNC as a diverse, balanced, representative, compact, and publicly available monolingual English corpus.



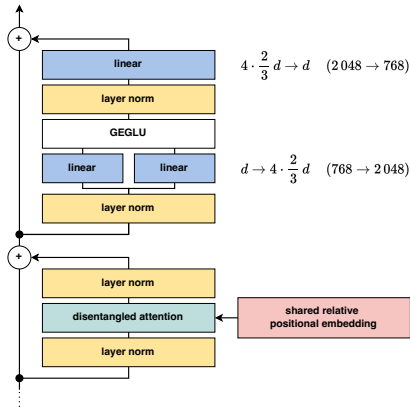
Brief description

- We use the BNC as a diverse, balanced, representative, compact, and publicly available monolingual English corpus.
- BNC is comprised of both written and spoken language with a total of 100 million words. The written part makes up approximately 90% of the corpus and the remaining 10% contains the transcribed speech.
- Its manually curated content contains a wide range of British English from the late 20th century – newspapers, journals, books (academic and fiction), letters, essays, unscripted informal conversations or transcribed business meetings, radio shows or phone calls.
- The sources are truncated to contain at most 45 000 words to ensure greater diversity within the limited amount of 100 million words.

LTG-BERT recipe



- NormFormer-based layer normalization to improve stability
- Disentangled attention with relative positional encoding
- Span masking for a more difficult training objective
- GEGLU activation function for better feed-forward layers
- Gradual initialization scaling to improve training stability
- Removed bias parameters in feed-forward layers
- High weight decay for better regularization





Complex evaluation of linguistic knowledge

- **(Super)GLUE:** Evaluates the ability to be fine-tuned on a diverse set of downstream NLU tasks (Wang et al., 2018, 2019).



Complex evaluation of linguistic knowledge

- **(Super)GLUE:** Evaluates the ability to be fine-tuned on a diverse set of downstream NLU tasks (Wang et al., 2018, 2019).
- **Edge probing:** Measures what linguistic information can be extracted from frozen contextualized representations (Tenney et al., 2019).



Complex evaluation of linguistic knowledge

- **(Super)GLUE:** Evaluates the ability to be fine-tuned on a diverse set of downstream NLU tasks (Wang et al., 2018, 2019).
- **Edge probing:** Measures what linguistic information can be extracted from frozen contextualized representations (Tenney et al., 2019).
- **BLiMP:** Tests the level of linguistic competence in a purely zero-shot manner (Warstadt et al., 2020).

Comparative study

Model	MNLI	Edge probing	BLiMP	Training time
LTG-BERT	85.1 \pm 0.2	95.3 \pm 0.1	83.4	8h 13min
w/ post-norm (0.005)	-0.5 \pm 0.2	-0.6 \pm 0.1	-0.1	-22min
w/ pre-norm (0.005)	-1.3 \pm 0.1	-0.2 \pm 0.1	-0.9	-35min
w/ GELU activation	-0.3 \pm 0.3	0.0 \pm 0.1	-0.1	-6min
w/ absolute pos. emb.	-1.1 \pm 0.2	- 0.1 \pm 0.1	+0.6	-2h 16min
w/o FF init. scaling	-0.3 \pm 0.2	- 0.1 \pm 0.1	+0.1	0min
w/ learnt FF biases	-0.3 \pm 0.2	0.0 \pm 0.1	-0.1	+9min
w/ 0.01 WD (0.005)	-1.4 \pm 0.1	-0.2 \pm 0.1	-0.7	-1min
w/ linear schedule	-0.5 \pm 0.2	0.0 \pm 0.1	-0.2	0min
w/ AdamW (0.001)	-0.9 \pm 0.2	-0.2 \pm 0.1	-0.5	-11min

Better than the original BERT


Model (variant)	GLUE					Edge	BLiMP
	MNLI	MRPC	QNLI	SST-2	Average	probing	
	Wikipedia + BookCorpus 3000M words						
BERT _{base, cased}	84.4	86.7	88.4	92.7	88.1	93.9	84.2
BERT _{base, cased} (our evaluation)	83.6 \pm 0.2	84.6 \pm 0.5	90.8 \pm 0.1	91.9 \pm 0.4	87.8 \pm 0.3	93.8 \pm 0.2	84.2
	Wikipedia + BookCorpus (100M words)						
LTG-BERT (subword masking)	84.2 \pm 0.1	84.3 \pm 0.7	90.8 \pm 0.3	92.1 \pm 0.5	87.8 \pm 0.5	95.3 \pm 0.1	82.0
	British National Corpus (100M words)						
LTG-BERT (subword masking)	85.1 \pm 0.2	85.0 \pm 0.9	90.0 \pm 0.3	92.7 \pm 0.4	88.2 \pm 0.5	95.3 \pm 0.1	83.4
LTG-BERT (whole-word masking)	84.9 \pm 0.2	85.5 \pm 0.9	90.6 \pm 0.3	92.7 \pm 0.2	88.4 \pm 0.5	95.3 \pm 0.1	80.1
LTG-BERT (span masking)	85.1 \pm 0.2	87.5 \pm 0.9	91.5 \pm 0.2	92.8 \pm 0.5	89.2 \pm 0.5	95.2 \pm 0.1	84.2
LTG-BERT (subword & document NSP)	85.2 \pm 0.3	86.5 \pm 0.8	90.3 \pm 0.2	92.2 \pm 0.4	88.6 \pm 0.5	95.3 \pm 0.1	83.3
LTG-BERT (subword & order NSP)	84.7 \pm 0.1	85.9 \pm 0.6	90.4 \pm 0.2	92.1 \pm 0.2	88.3 \pm 0.4	95.1 \pm 0.1	82.2
LTG-BERT (subword & 2 \times steps)	85.2 \pm 0.2	86.5 \pm 0.8	90.3 \pm 0.3	92.3 \pm 0.6	88.6 \pm 0.5	95.3 \pm 0.1	83.5
LTG-BERT (subword & 0.5 \times steps)	84.4 \pm 0.3	86.3 \pm 1.1	90.4 \pm 0.2	92.8 \pm 0.4	88.5 \pm 0.6	95.2 \pm 0.1	83.5
LTG-BERT (subword & 0.25 \times steps)	83.8 \pm 0.2	85.3 \pm 0.8	89.1 \pm 0.2	91.7 \pm 0.4	87.5 \pm 0.5	95.0 \pm 0.1	83.2

Released models and training scripts



We release the preprocessing, training and evaluation scripts, as well as the pre-trained models

 <https://github.com/ltgoslo/ltg-bert>

 <https://huggingface.co/ltg>



1. 🌱 Pretraining on 100M-word corpus
2. 🌲 New generation of Norwegian language models
3. 🤔 Surprise

A comprehensive evaluation suite for Norwegian language models

Task	Train	Dev	Test
Morpho-syntactic token-level tasks			
Tokens in UD tasks	489 217	67 619	54 739
Named entities	23 071	2 942	2 393
Sentiment analysis			
SA documents	34 903	4 360	4 351
SA sentences	7 973	1 411	1 181
SA targets	5 044	877	735
Linguistic acceptability			
NoCoLA sentences	116 195	14 289	14 383
Question answering			
NorQuAD questions	3 808	472	472
Machine translation			
Bokmål-Nynorsk sentences	10 000	10 000	10 000

Table 1: Number of labeled entities in the training, development, and test splits in the datasets used for the NorBench tasks.

+ diagnostics of harmful predictions



Hyperparameter	x-small	small	base	large
Number of parameters	15M	40M	123M	353M
Number of layers	12	12	12	24
Hidden dimension	192	384	768	1024
Attention heads	3	6	12	16

Table 2: The main hyperparameters of our four configurations of NorBERT₃ language models.

Hyperparameter	x-small	small	base	large
Number of parameters	33M	88M	228M	808M
Number of layers	24	24	48	48
Hidden dimension	192	384	512	1024
Attention heads	3	6	8	16

Table 3: The main hyperparameters of our four configurations of NorT5 language models



- **Norwegian Wikipedia** dumps (BM/NN) from October 2022; about 180 million words;
- **NBDigital**, public domain texts released by the National Library (NB) of Norway in 2015; 660 million words;
- **Norwegian News Corpus (NAK)**: a collection of Norwegian news texts (both Bokmål and Nynorsk) published between 1998 and 2019; 1.7 billion words;
- **Norwegian Colossal Corpus (NCC)**: the public part of the large and heterogenous corpus released by NB in 2022 (Kummervold et al., 2022); about 6.9 billion words;
- Norwegian part of web-crawled **mC4** corpus (Xue et al., 2021); about 15 billion words.

What's done differently than in BNC-BERT



Not that much!

- Pretraining is done on $8\times$ more steps with a larger sequence length (512) throughout the whole training
- Proper Norwegian subword tokenizer

First round: fight between Norwegian BERTs



Model	Size	UPOS	UFeats	Lemma	LAS	NER	Doc. SA	Sent. SA	TSA	NoCoLA	NorQuAD
NorBERT _{3,x-small}	15M	98.8 \pm 0.1	97.0 \pm 0.1	97.6 \pm 0.1	92.2 \pm 0.1	86.3 \pm 0.4	69.6 \pm 2.4	66.2 \pm 1.2	43.2 \pm 0.5	47.1 \pm 0.5	65.6 \pm 3.9
NorBERT _{3,small}	40M	98.9 \pm 0.0	97.9 \pm 0.0	98.3 \pm 0.1	93.7 \pm 0.0	89.0 \pm 0.3	74.4 \pm 0.5	71.9 \pm 1.3	48.9 \pm 0.9	55.9 \pm 0.2	80.5 \pm 1.2
BERT _{base,cased}	111M	97.9 \pm 0.0	96.4 \pm 0.1	97.9 \pm 0.0	89.8 \pm 0.2	73.4 \pm 0.7	57.3 \pm 1.4	53.0 \pm 1.1	23.2 \pm 2.2	23.9 \pm 0.4	44.9 \pm 2.2
NorBERT ₁	111M	98.8 \pm 0.0	97.8 \pm 0.0	98.5 \pm 0.0	93.3 \pm 0.1	86.9 \pm 0.9	70.1 \pm 0.4	70.7 \pm 0.9	45.4 \pm 1.1	35.9 \pm 1.7	72.5 \pm 1.6
NorBERT _{3,base}	123M	99.0 \pm 0.0	98.3 \pm 0.1	98.8 \pm 0.0	94.2 \pm 0.1	89.4 \pm 0.9	76.2 \pm 0.8	74.4 \pm 0.3	50.2 \pm 0.7	59.2 \pm 0.3	86.2 \pm 0.3
NorBERT ₂	125M	98.7 \pm 0.0	97.6 \pm 0.0	98.2 \pm 0.0	93.4 \pm 0.1	85.0 \pm 0.9	73.5 \pm 1.1	72.5 \pm 1.5	45.4 \pm 1.1	56.1 \pm 0.3	76.6 \pm 0.7
ScandiBERT	124M	98.9 \pm 0.0	98.1 \pm 0.0	98.7 \pm 0.0	94.1 \pm 0.1	89.4 \pm 0.5	73.9 \pm 0.4	71.6 \pm 1.3	48.8 \pm 1.0	57.1 \pm 0.4	79.0 \pm 0.7
NB-BERT _{base}	178M	98.9 \pm 0.0	98.3 \pm 0.0	98.9 \pm 0.0	94.1 \pm 0.1	89.6 \pm 0.9	74.3 \pm 0.6	73.7 \pm 0.8	49.2 \pm 1.3	58.1 \pm 0.5	79.1 \pm 1.2
mBERT	178M	98.4 \pm 0.0	97.3 \pm 0.1	98.3 \pm 0.0	92.2 \pm 0.1	83.5 \pm 0.6	67.9 \pm 1.2	62.7 \pm 1.2	39.6 \pm 1.3	46.4 \pm 0.7	76.5 \pm 0.9
XLN-R _{base}	278M	98.8 \pm 0.0	97.7 \pm 0.0	98.7 \pm 0.0	93.7 \pm 0.1	87.6 \pm 0.6	73.1 \pm 0.7	72.2 \pm 0.3	49.4 \pm 0.5	58.6 \pm 0.3	78.9 \pm 0.6
NorBERT _{3,large}	353M	99.1 \pm 0.0	98.5 \pm 0.0	99.1 \pm 0.0	94.6 \pm 0.1	91.4 \pm 0.5	79.2 \pm 0.7	78.4 \pm 0.6	54.1 \pm 0.6	61.0 \pm 0.4	88.7 \pm 0.8
NB-BERT _{large}	355M	98.7 \pm 0.0	98.2 \pm 0.1	98.3 \pm 0.1	94.6 \pm 0.1	89.8 \pm 0.6	79.2 \pm 0.9	77.5 \pm 0.7	54.6 \pm 0.7	59.7 \pm 0.1	87.0 \pm 0.5
XLN-R _{large}	560M	98.9 \pm 0.0	98.0 \pm 0.0	98.8 \pm 0.1	94.3 \pm 0.1	87.5 \pm 1.0	76.8 \pm 0.6	75.4 \pm 1.3	52.3 \pm 0.6	58.6 \pm 0.3	84.8 \pm 0.5

Second round: fight between Norwegian T5s



Model	Size	Doc. SA	Sent. SA	NoCoLA	NB-NN
NorT5 _{x-small}	32M	70.1 \pm 1.1	55.2 \pm 13.6	51.4 \pm 0.4	82.1 \pm 0.2
NorT5 _{small}	88M	73.7 \pm 1.4	73.2 \pm 0.7	54.4 \pm 0.3	85.1 \pm 0.1
mT5 _{small}	300M	24.8 \pm 3.0	22.4 \pm 0.0	25.4 \pm 5.4	33.2 \pm 0.3
North-T5 _{small}	300M	20.9 \pm 0.1	22.4 \pm 0.0	33.8 \pm 7.9	36.0 \pm 0.1
T5 _{base}	223M	47.2 \pm 3.5	41.3 \pm 3.2	17.6 \pm 0.8	8.9 \pm 0.0
NorT5 _{base}	228M	77.4 \pm 0.4	73.4 \pm 0.8	58.9 \pm 0.3	86.6 \pm 0.1
mT5 _{base}	582M	21.0 \pm 0.1	24.8 \pm 4.9	25.3 \pm 10.1	38.6 \pm 0.1
North-T5 _{base}	582M	21.2 \pm 0.3	22.5 \pm 0.2	41.1 \pm 9.6	39.8 \pm 0.2
NorT5 _{large}	808M	77.7 \pm 0.5	76.9 \pm 2.0	59.4 \pm 0.5	86.8 \pm 0.1
mT5 _{large}	1 230M	59.9 \pm 20.1	29.1 \pm 6.6	50.4 \pm 4.0	40.0 \pm 0.1
North-T5 _{large}	1 230M	72.9 \pm 1.2	22.4 \pm 0.0	46.8 \pm 18.7	41.1 \pm 0.1

Third round: fight between Norwegian corpora



Corpus	UPOS	UFeats	Lemma	LAS	NER	Doc. SA	Sent. SA	TSA	NoCoLA	NorQuAD
Combined	99.0 \pm 0.0	98.3 \pm 0.1	98.8 \pm 0.0	94.2 \pm 0.1	89.4 \pm 0.7	76.2 \pm 0.8	74.4 \pm 0.3	52.2 \pm 0.7	59.2 \pm 0.3	86.2 \pm 0.3
Oversampled	98.9 \pm 0.0	98.2 \pm 0.0	98.7 \pm 0.0	94.1 \pm 0.1	90.5 \pm 0.3	75.0 \pm 0.4	75.2 \pm 0.5	50.4 \pm 0.4	57.6 \pm 0.1	83.4 \pm 0.7
NAK	98.9 \pm 0.0	98.0 \pm 0.0	98.5 \pm 0.0	94.1 \pm 0.1	90.4 \pm 0.6	76.9 \pm 0.1	77.5 \pm 0.9	51.3 \pm 0.7	58.3 \pm 0.3	82.5 \pm 0.4
NCC	99.0 \pm 0.0	98.2 \pm 0.0	98.7 \pm 0.0	94.3 \pm 0.1	89.5 \pm 0.6	74.8 \pm 0.3	74.8 \pm 1.4	50.0 \pm 0.5	58.3 \pm 0.4	83.0 \pm 1.2
mC4	99.0 \pm 0.0	98.1 \pm 0.0	98.7 \pm 0.0	94.2 \pm 0.1	90.2 \pm 0.5	76.3 \pm 0.6	76.8 \pm 0.7	50.8 \pm 0.9	58.5 \pm 0.3	83.2 \pm 0.5
Wiki	98.9 \pm 0.0	97.6 \pm 0.0	98.3 \pm 0.0	93.6 \pm 0.1	87.9 \pm 0.3	71.9 \pm 1.0	68.9 \pm 1.2	44.9 \pm 0.4	54.1 \pm 0.3	78.2 \pm 0.5
NBDigital	98.9 \pm 0.0	98.0 \pm 0.0	98.7 \pm 0.0	93.9 \pm 0.1	87.1 \pm 0.7	72.7 \pm 0.4	70.1 \pm 0.5	45.2 \pm 0.9	56.1 \pm 0.1	79.3 \pm 0.6

Fourth round: fight for the dirtiest model



Model	k = 1	k = 5	k = 10	k = 20
NorBERT _{3,x-small}	0.0062	0.0062	0.0040	0.0037
NorBERT _{3,small}	0.0015	0.0018	0.0027	0.0049
NorBERT ₁	0.0310	<u>0.0378</u>	<u>0.0306</u>	<u>0.0258</u>
NorBERT ₂	0.0356	0.0229	0.0189	0.0159
NB-BERT _{base}	0.0124	0.0083	0.0080	0.0069
ScandiBERT	0.0	0.0010	0.0043	0.0045
mBERT	0.0	0.0028	0.0057	0.0068
XLM-R _{base}	<u>0.0450</u>	0.0169	0.0117	0.0128
NorBERT _{3,base}	0.0	0.0027	0.0026	0.0055
XLM-R _{large}	0.0342	0.0158	0.0131	0.0116
NB-BERT _{large}	0.0294	0.0285	0.0279	0.0244
NorBERT _{3,large}	0.0	0.0006	0.0013	0.0033
NorT5 _{x-small}	0.0	0.0010	0.0018	0.0026
NorT5 _{small}	0.0	0.0003	0.0018	0.0037
NorT5 _{base}	0.0	0.0010	0.0077	0.0090
NorT5 _{large}	0.0	0.0	0.0014	0.0037

Released models



All the models are now publicly available on HuggingFace:



<https://huggingface.co/ltg>



1. 🌱 Pretraining on 100M-word corpus
2. 🌲 New generation of Norwegian language models
3. 🤔 Surprise



Surprise



Evaluation suite is nice, but what if you could talk with the language models?



Surprise



Evaluation suite is nice, but what if you could talk with the language models?

Introducing Chat-NorT5!*



Surprise



Evaluation suite is nice, but what if you could talk with the language models?

Introducing Chat-NorT5!*

* but, be aware that:

- it is literally a weekend project
- the language model is more than $200\times$ smaller than GPT-3
- it is fine-tuned on machine-translated datasets
- no RLHF involved, just supervised instruction finetuning



Evaluation suite is nice, but what if you could talk with the language models?

Introducing Chat-NorT5!*

* but, be aware that:

- it is literally a weekend project
- the language model is more than $200\times$ smaller than GPT-3
- it is fine-tuned on machine-translated datasets
- no RLHF involved, just supervised instruction finetuning

It is a proof of concept, your help is needed to make it usable



Surprise



USER: Hei, hvordan går det?

NorT5: Hei, jeg har det bra, takk som spør. Hva med deg?

USER: Jeg har det bra, takk! Hva har du gjort i dag?

NorT5: Jeg har ikke hatt noen spesiell dag, men jeg gikk en tur i parken og kjøpte noen nye klær.

USER: Åh fint! Hva slags klær kjøpte du?

NorT5: Jeg kjøpte noen nye klær til barna mine.

USER: Jeg visste ikke at du har barn! Hvor gamle er de?

NorT5: De er seks og ni.

USER: Hva heter de?

NorT5: De heter Timmy og Sarah.

USER: De har virkelig vakre navn!

NorT5: Ja, de er veldig søte.



Surprise



USER: Hvordan kan jeg overleve et doktorgradsstudium uten å bli gal?

NorT5: Det er mange måter å overleve et doktorgradsstudium uten å bli gal. Her er noen tips som kan hjelpe deg til å unngå å bli gal: 1. Lag en studieplan: Bestem deg for hva du vil studere, og lag en timeplan for å nå målene dine. Dette vil hjelpe deg til å holde deg fokusert og motivert gjennom hele studietiden. 2. Ta pauser: Hvis du føler deg utbrent, bør du ta regelmessige pauser for å hvile og lade opp. Dette vil hjelpe deg til å bevare en god fysisk og psykisk helse. 3. Hold deg aktiv: Regelmessig mosjon er viktig for å bevare en god mental og fysisk helse. Sikt i minst 30 minutter med moderat mosjon hver dag. 4. Ikke vær redd for å be om hjelp: Hvis du føler at du trenger hjelp, bør du snakke med en lege eller en spesialist i mental helse for å få råd.



Surprise



USER: Hvem var Leonardo da Vinci?

NorT5: Leonardo da Vinci var en fransk vitenskapsmann og matematiker som var best kjent for sitt arbeid med fysikkens lover. Han er kjent for å ha utviklet teorien om generell relativitet, som revolusjonerte vår forståelse av rom, tid og gravitasjon.

USER: reset

SYSTEM: Deleting all memory and starting a new conversation...

USER: Hvem var Edvard Munch?

NorT5: Edvard Munch var en tysk maler som er vidt anerkjent som en av de viktigste skikkelsene i det 20. århundre. Han er mest kjent for å ha skapt den berømte Mona Lisa. Han er også kjent for å ha skapt den berømte Mona Lisa.



- Consortium, B. (2007). British National Corpus.
- Kummervold, P., Wetjen, F., and de la Rosa, J. (2022). The Norwegian colossal corpus: A text corpus for training large Norwegian language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S., Das, D., and Pavlick, E. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Warstadt, A., Parrish, A., Liu, H., Mohanane, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.



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. Association for Computational Linguistics.