# NorBERT and other puffins

Andrey Kutuzov
Language Technology Group

25 January 2021

# Contents


1. Large-scale LMs for Norwegian
2. NorELMo
3. NorBERT
4. Evaluation
5. Thanks!

# Large-scale LMs for Norwegian

**LTG Oslo**
@ltgoslo

Today LTG Oslo releases large-scale language models for Norwegian
This includes NorBERT, the 1st fully functional Norwegian BERT, trained on Wikipedia and news (Norsk Aviskorpus).
Also available via #HuggingFace as ltgoslo/norbert

Details and download:
norlm.nlpl.eu

2:08 pm · 13 Jan 2021 · Twitter Web App

# Large-scale LMs for Norwegian

## But it's not just BERT

- `http://norlm.nlpl.eu`
- NorLM is an emerging collection of large-scale contextualized language models for Norwegian.
- Joint initiative:
  - EOSC-Nordic (European Open Science Cloud)
  - SANT (Sentiment Analysis for Norwegian)
  - coordinated by the Language Technology Group (LTG) at the University of Oslo

# Large-scale LMs for Norwegian

## But it's not just BERT

- `http://norlm.nlpl.eu`
- NorLM is an emerging collection of large-scale contextualized language models for Norwegian.
- Joint initiative:
  - EOSC-Nordic (European Open Science Cloud)
  - SANT (Sentiment Analysis for Norwegian)
  - coordinated by the Language Technology Group (LTG) at the University of Oslo
- Aim: to provide models and supporting tools for NLP researchers and developers for Norwegian.
- All the models are publicly available for download from the NLPL Vectors Repository:
- `http://vectors.nlpl.eu/repository`

# Large-scale LMs for Norwegian

## But it's not just BERT

- `http://norlm.nlpl.eu`
- NorLM is an emerging collection of large-scale contextualized language models for Norwegian.
- Joint initiative:
  - EOSC-Nordic (European Open Science Cloud)
  - SANT (Sentiment Analysis for Norwegian)
  - coordinated by the Language Technology Group (LTG) at the University of Oslo
- Aim: to provide models and supporting tools for NLP researchers and developers for Norwegian.
- All the models are publicly available for download from the NLPL Vectors Repository:
- `http://vectors.nlpl.eu/repository`

We are ever thankful to the Norwegian national supercomputing services operated by UNINETT Sigma2, the National Infrastructure for High Performance Computing and Data Storage in Norway.

# Large-scale LMs for Norwegian

## Model types

- Static embeddings [Mikolov et al., 2013]
- LSTM-based contextualized embeddings (ELMo) [Peters et al., 2018]
- Transformer-based contextualized embeddings (BERT) [Devlin et al., 2019]

All the models are accessible from Saga: `/cluster/shared/nlpl/data/vectors/latest`

# Large-scale LMs for Norwegian

## Model types

- Static embeddings [Mikolov et al., 2013]
- LSTM-based contextualized embeddings (ELMo) [Peters et al., 2018]
- Transformer-based contextualized embeddings (BERT) [Devlin et al., 2019]

# Contents


1. Large-scale LMs for Norwegian
2. NorELMo
3. NorBERT
4. Evaluation
5. Thanks!

# NorELMo

- ELMo: the first well-known contextualized embedding architecture.
- Two layers of bidirectional LSTMs.
- Best Paper award at the NAACL 2018 [Peters et al., 2018]
- Memory requirements are much more relaxed than transformers
- In many cases, delivers comparable performance.

# NorELMo

## NLPL Vector repository features 2 ELMo models for Norwegian

- 210: trained on lemmatized Norwegian Bokmål Wikipedia Dump of September 2020
- 211: trained on tokenized Norwegian Bokmål Wikipedia Dump of September 2020
- First full-fledged ELMo models for Norwegian (ElmoForManyLangs project used very small corpora samples)

## Training corpus

- 160 million words
- tokenized and lemmatized with UDPipe
- 100 000 most frequent words used as the ELMo vocabulary

## Training

- 3 epochs with batch size 192
- LSTM dimensionality 2048 (default 4096)

## NorELMO usage

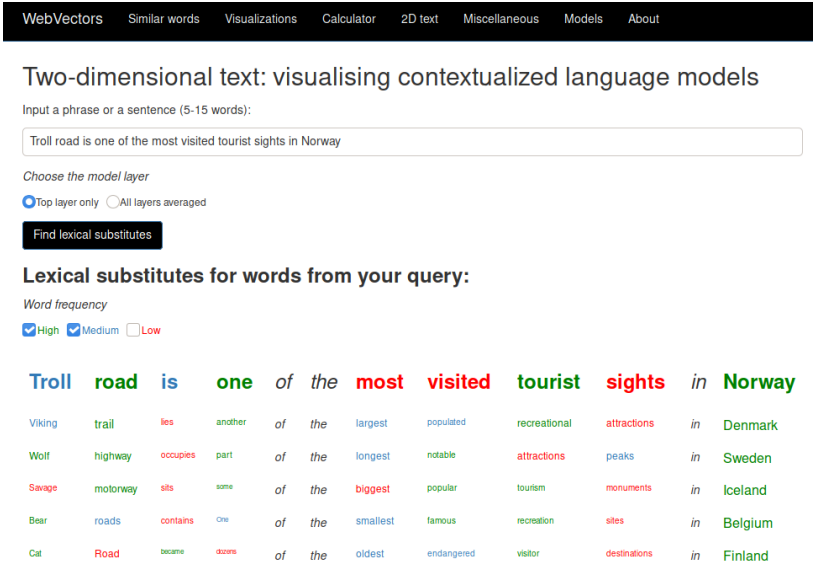
- Models released as TensorFlow checkpoints and HDF5 files
- fully self-contained
- the tool of choice is `Simple_elmo`:
  - `https://pypi.org/project/simple-elmo/`
  - Depends on TensorFlow (works with either TF1 or TF2)
  - A modernized code from the ELMo authors, made more convenient for everyday usage.

## NorELMO usage

- Models released as TensorFlow checkpoints and HDF5 files
- fully self-contained
- the tool of choice is `Simple_elmo`:
  - `https://pypi.org/project/simple-elmo/`
  - Depends on TensorFlow (works with either TF1 or TF2)
  - A modernized code from the ELMo authors, made more convenient for everyday usage.
- On Saga:
  - `module use -a /cluster/projects/nn9851k/software/easybuild/install/modules/all/`
  - `module load NLPL-simple_elmo/0.6.0-gomkl-2019b-Python-3.7.4`
  - `from simple_elmo import ElmoModel`
  - `model = ElmoModel()`
  - `model.load('/cluster/shared/nlpl/data/vectors/latest/211.zip')`
  - `model.get_elmo_vectors([['Han', 'vil', 'sove']])`

# NorELMo

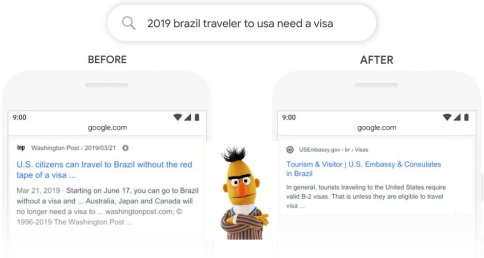
http://vectors.nlpl.eu/explore/embeddings/en/contextual/



# Contents


1. Large-scale LMs for Norwegian
2. NorELMo
3. NorBERT
4. Evaluation
5. Thanks!

# NorBERT




- Bidirectional Encoder Representations from Transformers
- Best Paper award at the NAACL 2019 [Devlin et al., 2019]
- Bread and butter of modern NLP
- The only BERT for Norwegian used to be the Google's Multilingual BERT (mBERT) :-(

# NorBERT

## We trained NorBERT from scratch on Norwegian data

- The training largely followed the trail laid by FinBERT (`https://github.com/TurkuNLP/FinBERT`)
- Important: a custom 30 000 SentencePiece vocabulary (cased with diacritics
- *'Denne gjengen håper at de sammen skal bidra til å gi kvinnefotballen i Kristiansand et lenge etterlengtet løft.'*

# NorBERT

## We trained NorBERT from scratch on Norwegian data

- The training largely followed the trail laid by FinBERT (`https://github.com/TurkuNLP/FinBERT`)
- Important: a custom 30 000 SentencePiece vocabulary (cased with diacritics
- *'Denne gjengen håper at de sammen skal bidra til å gi kvinnefotballen i Kristiansand et lenge etterlengtet løft.'*
  - mBERT: 'Denne g ##jeng ##en h ##å ##per at de sammen skal bid ##ra til å gi k ##vinne ##fo ##t ##ball ##en i Kristiansand et lenge etter ##len ##gte ##t l ##ø ##ft .'

# NorBERT

## We trained NorBERT from scratch on Norwegian data

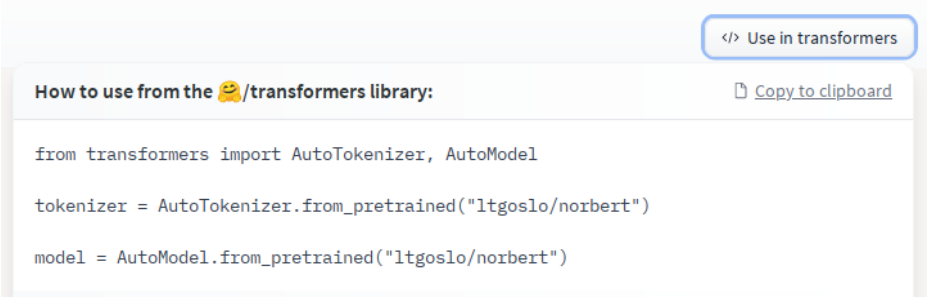
- The training largely followed the trail laid by FinBERT (`https://github.com/TurkuNLP/FinBERT`)
- Important: a custom 30 000 SentencePiece vocabulary (cased with diacritics
- *'Denne gjengen håper at de sammen skal bidra til å gi kvinnefotballen i Kristiansand et lenge etterlengtet løft.'*
  - mBERT: 'Denne g ##jeng ##en h ##å ##per at de sammen skal bid ##ra til å gi k ##vinne ##fo ##t ##ball ##en i Kristiansand et lenge etter ##len ##gte ##t l ##ø ##ft .'
  - NorBERT: 'Denne gjengen håper at de sammen skal bidra til å gi kvinne ##fotball ##en i Kristiansand et lenge etterl ##engt ##et løft .'

# NorBERT

## We trained NorBERT from scratch on Norwegian data

- The training largely followed the trail laid by FinBERT (`https://github.com/TurkuNLP/FinBERT`)
- Important: a custom 30 000 SentencePiece vocabulary (cased with diacritics
- *'Denne gjengen håper at de sammen skal bidra til å gi kvinnefotballen i Kristiansand et lenge etterlengtet løft.'*
  - mBERT: 'Denne g ##jeng ##en h ##å ##per at de sammen skal bid ##ra til å gi k ##vinne ##fo ##t ##ball ##en i Kristiansand et lenge etter ##len ##gte ##t l ##ø ##ft .'
  - NorBERT: 'Denne gjengen håper at de sammen skal bidra til å gi kvinne ##fotball ##en i Kristiansand et lenge etterl ##engt ##et løft .'
- Where?
  - `http://vectors.nlpl.eu/repository/20/215.zip`
  - `https://huggingface.co/ltgoslo/norbert`

# NorBERT




...or simply load from `/cluster/shared/nlpl/data/vectors/latest/215.zip` on Saga.

# NorBERT

## Training Corpus

1. Norsk Aviskorpus (NAK); 1.7 billion words;
2. Bokmål Wikipedia; 160 million words;
3. Nynorsk Wikipedia; 40 million words;

2 billion word tokens in 200 million sentences.
Sentence-segmented using Stanza.

# NorBERT

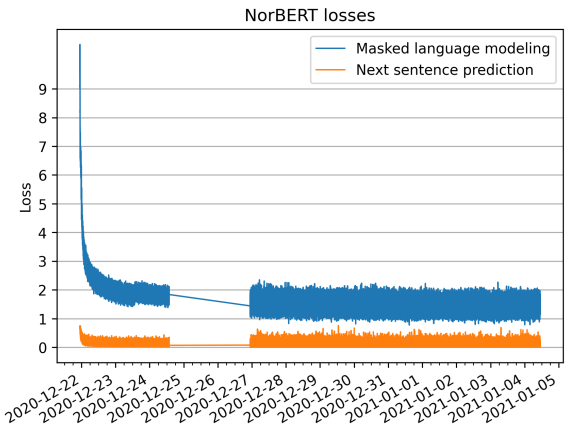
## How it was done?

- Similar to Bert-Base Cased (12 layers, hidden size 768)
- Trained on Saga for about 3 weeks
- 4 compute nodes, 16 NVIDIA P100 GPUs
- BERT implementation by NVIDIA: multi-node and multi-GPU training
- `http://wiki.nlpl.eu/index.php/Eosc/pretraining/nvidia`
- Related code: `https://github.com/ltgoslo/NorBERT`

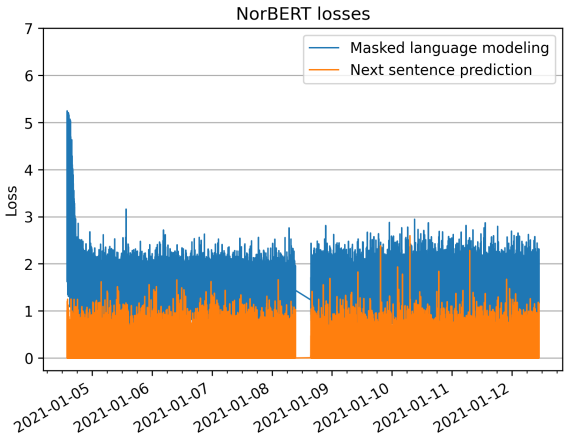
## Training workflow

- Phase 1
  - training with maximum sequence length of 128
  - batch size 48 and global batch size 48*16=768
  - 265 000 training steps = 1 epoch
  - We have done 3 epochs: 795 000 training steps.
- Phase 2
  - training with maximum sequence length of 512
  - batch size 8 and global batch size 8*16=128
  - aimed at 1/9 of the number of sentences seen during Phase 1
  - 68 million sentences: 531 000 training steps.

# NorBERT



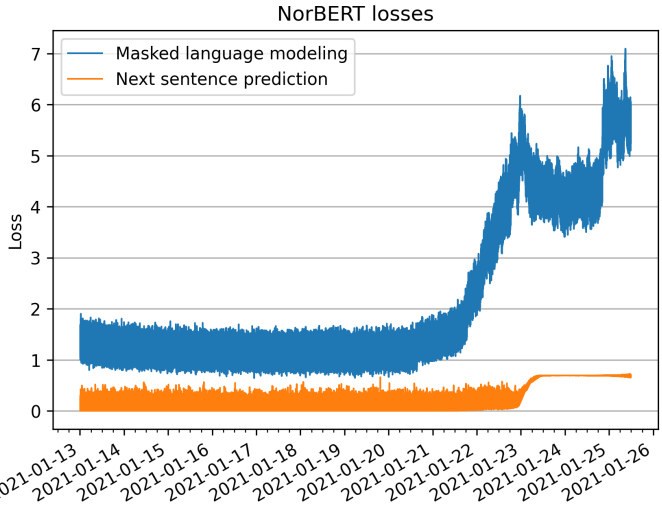

Phase 1 training




Phase 2 training

# NorBERT

Currently a 10-epoch model is being trained. But weird stuff is going on:



# Contents


1. Large-scale LMs for Norwegian
2. NorELMo
3. NorBERT
4. Evaluation
5. Thanks!

# Evaluation

Let's compare NorBERT to mBERT and to the recently released NB-BERT-Base from the National Library NoTraM project (`https://github.com/NBAiLab/notram`).
NB-BERT-Base is trained on 10 times more data than NorBERT!

# Evaluation

Let's compare NorBERT to mBERT and to the recently released NB-BERT-Base from the National Library NoTraM project (`https://github.com/NBAiLab/notram`).
NB-BERT-Base is trained on 10 times more data than NorBERT!

## Part-of-speech tagging (Bokmål Universal Dependencies 2.7)

| Metrics | mBERT | NorBERT | NB-BERT-Base |
|---|---|---|---|
| Accuracy | 97.7 | 98.4 | **98.5** |

## Sentence-level binary sentiment classification (NoReC_fine)

| Metrics | mBERT | NorBERT | NB-BERT-Base |
|---|---|---|---|
| F1 score | 67.0 | **81.8** | 80.5 |

Not bad at all!

# Contents

1. Large-scale LMs for Norwegian
2. NorELMo
3. NorBERT
4. Evaluation
5. Thanks!

# Thanks!

https://huggingface.co/ltgoslo/norbert

# References I


- 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. Association for Computational Linguistics.

- Mikolov, T., Yih, W.-t., and Zweig, G. (2013).
  Linguistic regularities in continuous space word representations.
  In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

# References II


Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.