OCR Models for Sámi Languages

This is a collection of models for OCR (optical character recognition) of Sámi languages. These can be used to recognize text in images of printed text (scanned books, magazines, etc.) in North Sámi, South Sámi, Lule Sámi, and Inari Sámi.

You can read more detailed information about the training and evaluation of the models in the article "Comparative analysis of optical character recognition methods for Sámi texts from the National Library of Norway". Link to preprint: https://arxiv.org/abs/2501.07300

The collection consists of three different types of models: Transkribus models, Tesseract models, and TrOCR models.

Transkribus Models

The Transkribus models are available in the application Transkribus.
Transkribus is a tool where you can use models to recognize text in images and correct any prediction errors.
It may cost money to use Transkribus, depending on the amount of data to be transcribed.

The following models are available:

- SamiskOCR_smi_ub (model-id 181605): a model trained exclusively on manually annotated Sámi data
- SamiskOCR_smi (model-id 181725): print M1 base model (model-id 39995) fine-tuned on manually annotated Sámi data
- SamiskOCR_smi_nor (model-id 182005): print M1 base model fine-tuned on manually annotated Sámi and Norwegian data
- SamiskOCR_smi_smipred (model-id 192137): print M1 base model fine-tuned on manually annotated and automatically transcribed Sámi data
- SamiskOCR_alt (model-id 179305): print M1 base model fine-tuned on manually annotated and automatically transcribed Sámi data, and manually annotated Norwegian data

The latter model achieved the best results on our test dataset.

Tesseract Models

Tesseract is an OCR tool where you can run models locally.
See the installation guide to install Tesseract.
Once you have installed Tesseract, you can use the .traineddata model files in `tesseract_models/` to run OCR on your images.

The following models are available:

- ub_smi: a model trained exclusively on manually annotated Sámi data
- smi: the Norwegian base model fine-tuned on manually annotated Sámi data
- smi_nor: the Norwegian base model fine-tuned on manually annotated Sámi and Norwegian data
- smi_pred: the Norwegian base model fine-tuned on manually annotated and automatically transcribed Sámi data

- smi_nor_pred: the Norwegian base model fine-tuned on manually annotated and automatically transcribed Sámi data, and manually annotated Norwegian data
- synth_base: the Norwegian base model fine-tuned on synthetic* Sámi data
- sb_smi: synth_base fine-tuned on manually annotated Sámi data
- sb_smi_nor_pred: synth_base fine-tuned on manually annotated and automatically transcribed Sámi data, and manually annotated Norwegian data

The latter model achieved the best results (of the Tesseract models) on our test dataset.

* synthetic here means that we have real Sámi text, which we have created images of, to resemble scanned text. Unlike the manually transcribed data, which are books and newspapers that are scanned and then manually transcribed.

Tesseract tips:

- You need to move the .traineddata files to the tessdata area on your PC
- To run OCR with the desired model, you must specify the model name with -l, e.g., `tesseract filename.jpg outfile -l smi_nor_pred`. Then `outfile.txt` will contain the text that the model found in the image
- Use --psm 7 if you have an image of a line of text, not a whole page. See more info here

TrOCR Models

TrOCR is a transformers-based model architecture for text recognition.
All our TrOCR models are available from our huggingface collection, and in `trocr_models/` in this directory
You can use the models with the transformers library

The following models are available:

- trocr_smi: TrOCR-printed base model fine-tuned on manually annotated Sámi data
- trocr_smi_nor: TrOCR-printed base model fine-tuned on manually annotated Sámi and Norwegian data
- trocr_smi_pred: TrOCR-printed base model fine-tuned on manually annotated and automatically transcribed Sámi data
- trocr_smi_nor_pred: TrOCR-printed base model fine-tuned on manually annotated and automatically transcribed Sámi data, and manually annotated Norwegian data
- trocr_smi_synth: TrOCR-printed base model fine-tuned on synthetic* Sámi data, and then on manually annotated Sámi data
- trocr_smi_pred_synth: TrOCR-printed base model fine-tuned on synthetic Sámi data, and then fine-tuned on manually annotated and automatically transcribed Sámi data
- trocr_smi_nor_pred_synth: TrOCR-printed base model fine-tuned on synthetic Sámi data, and then fine-tuned on manually annotated and automatically transcribed Sámi data, and manually annotated Norwegian

trocr_smi_pred_synth is the model that achieved the best results (of the TrOCR models) on our test dataset.

* synthetic here means that we have real Sámi text, which we have created images of, to resemble scanned text. Unlike the manually transcribed data, which are books and newspapers that are scanned and then

manually transcribed.

The models only work with images of lines of text. If you have images of entire pages of text, you must segment the text into lines first to benefit from this model.

Usage example with python

```python
from transformers import TrOCRProcessor, VisionEncoderDecoderModel
from PIL import Image

# EITHER
# download model from hf hub
processor = TrOCRProcessor.from_pretrained("Sprakbanken/trocr_smi_pred_synth")
model = VisionEncoderDecoderModel.from_pretrained("Sprakbanken/trocr_smi_pred_synth")

# OR
# import local model
processor = TrOCRProcessor.from_pretrained("<relative_path_to>/trocr_models/trocr_smi_pred_synth")
model = VisionEncoderDecoderModel.from_pretrained("<relative_path_to>/trocr_models/trocr_smi_pred_synth")

image = Image.open("path_to_image.jpg").convert("RGB")

pixel_values = processor(image, return_tensors="pt").pixel_values
generated_ids = model.generate(pixel_values)

generated_text = processor.batch_decode(generated_ids, skip_special_tokens=True)[0]
```

License, use and citation

The models in the collection are shared under CC BY 4.0 (link: https://creativecommons.org/licenses/by/4.0/deed.en)

You can freely reuse and modify the models, but must cite the article:
"Enstad T, Trosterud T, Røsok MI, Beyer Y, Roald M. Comparative analysis of optical character recognition methods for Sámi texts from the National Library of Norway. Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa) 2025."