DASC7606
2025/11/06

Assignment II: Machine Translation

Weijian Deng (kendwj@hku.hk)

Shu Chen (schen59@hku.hk)

Find it on Github: https://github.com/hkukend/DASC7606A-B-A2

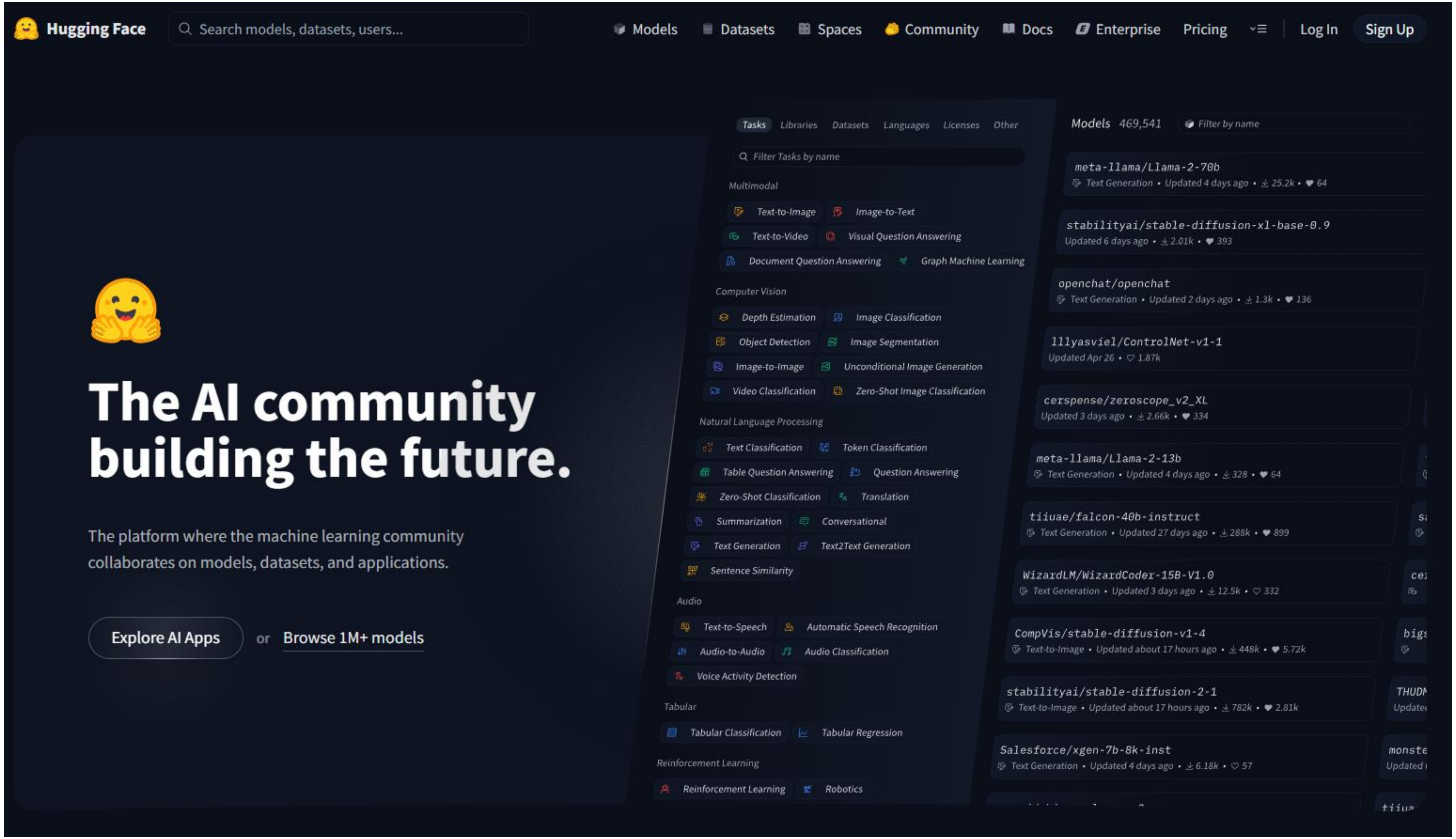
Optional Assignment:

- **This assignment is optional** due to the coming final exams
- **The final assignment grades** will be the maximum between Assignment 1 and 2

Objectives:

- **Use the Hugging Face Hub for datasets:** discover, download, and prepare translation datasets
- **Load pretrained models and tokenizers** from the community
- **Build training pipelines using** transformers, datasets, accelerate, and evaluate
- **Fine-tune a translation model** end-to-end for a chosen language pair (zh-sim->en)
- **Evaluate translation quality** using BLEU (via sacrebleu) and report results
- **Develop debugging skills** for identifying and fixing common deep learning issues in HF version

HuggingFace🤗:



The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Multimodal

- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Sentence Similarity

Audio

- Text-to-Speech
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

Tabular

- Tabular Classification
- Tabular Regression

Reinforcement Learning

- Reinforcement Learning
- Robotics

Models 469,541

meta-llama/Llama-2-70b
Text Generation • Updated 4 days ago • 25.2k • ♥ 64

stabilityai/stable-diffusion-xl-base-0.9
Updated 6 days ago • 2.01k • ♥ 393

openchat/openchat
Text Generation • Updated 2 days ago • 1.3k • ♥ 136

lllyasviel/ControlNet-v1-1
Updated Apr 26 • ♡ 1.87k

cerspense/zeroscope_v2_XL
Updated 3 days ago • 2.66k • ♥ 334

meta-llama/Llama-2-13b
Text Generation • Updated 4 days ago • 328 • ♥ 64

tiiuae/falcon-40b-instruct
Text Generation • Updated 27 days ago • 288k • ♥ 899

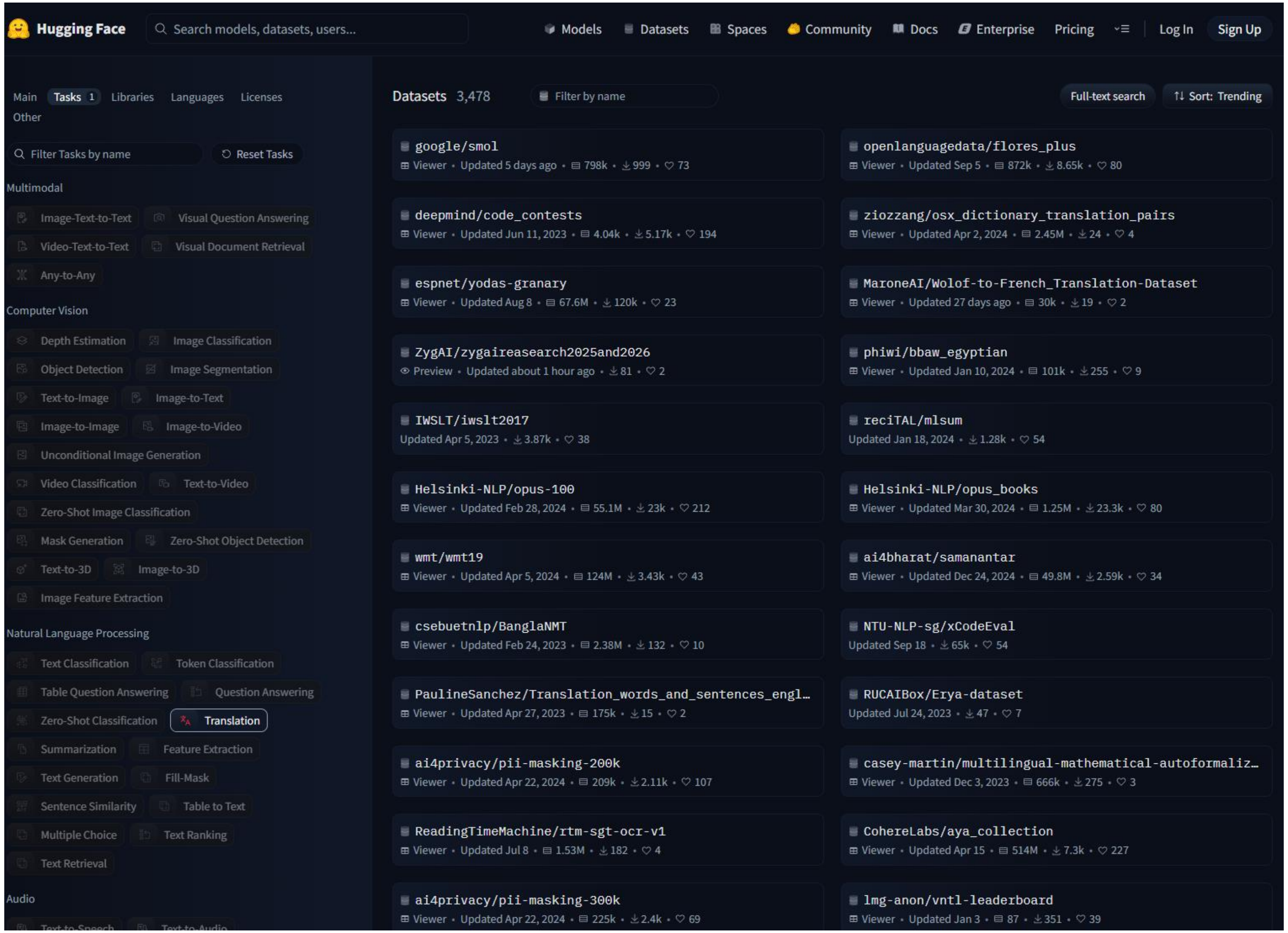
WizardLM/WizardCoder-15B-V1.0
Text Generation • Updated 3 days ago • 12.5k • ♡ 332

CompVis/stable-diffusion-v1-4
Text-to-Image • Updated about 17 hours ago • 448k • ♥ 5.72k

stabilityai/stable-diffusion-2-1
Text-to-Image • Updated about 17 hours ago • 782k • ♥ 2.81k

Salesforce/xgen-7b-8k-inst
Text Generation • Updated 4 days ago • 6.18k • ♡ 57

HuggingFace 🤗 Datasets:



Datasets 3,478

google/smol
⊞ Viewer · Updated 5 days ago · ▤ 798k · ↓999 · ♡ 73

openlanguagedata/flores_plus
⊞ Viewer · Updated Sep 5 · ▤ 872k · ↓8.65k · ♡ 80

deepmind/code_contests
⊞ Viewer · Updated Jun 11, 2023 · ▤ 4.04k · ↓5.17k · ♡ 194

ziozzang/osx_dictionary_translation_pairs
⊞ Viewer · Updated Apr 2, 2024 · ▤ 2.45M · ↓24 · ♡ 4

espnet/yodas-granary
⊞ Viewer · Updated Aug 8 · ▤ 67.6M · ↓120k · ♡ 23

MaroneAI/Wolof-to-French_Translation-Dataset
⊞ Viewer · Updated 27 days ago · ▤ 30k · ↓19 · ♡ 2

ZygAI/zygaireasearch2025and2026
Preview · Updated about 1 hour ago · ↓81 · ♡ 2

phiwi/bbaw_egyptian
⊞ Viewer · Updated Jan 10, 2024 · ▤ 101k · ↓255 · ♡ 9

IWSLT/iwslt2017
Updated Apr 5, 2023 · ↓3.87k · ♡ 38

reciTAL/mlsum
Updated Jan 18, 2024 · ↓1.28k · ♡ 54

Helsinki-NLP/opus-100
⊞ Viewer · Updated Feb 28, 2024 · ▤ 55.1M · ↓23k · ♡ 212

Helsinki-NLP/opus_books
⊞ Viewer · Updated Mar 30, 2024 · ▤ 1.25M · ↓23.3k · ♡ 80

wmt/wmt19
⊞ Viewer · Updated Apr 5, 2024 · ▤ 124M · ↓3.43k · ♡ 43

ai4bharat/samanantar
⊞ Viewer · Updated Dec 24, 2024 · ▤ 49.8M · ↓2.59k · ♡ 34

csebuetnlp/BanglaNMT
⊞ Viewer · Updated Feb 24, 2023 · ▤ 2.38M · ↓132 · ♡ 10

NTU-NLP-sg/xCodeEval
Updated Sep 18 · ↓65k · ♡ 54

PaulineSanchez/Translation_words_and_sentences_engl…
⊞ Viewer · Updated Apr 27, 2023 · ▤ 175k · ↓15 · ♡ 2

RUCAIBox/Erya-dataset
Updated Jul 24, 2023 · ↓47 · ♡ 7

ai4privacy/pii-masking-200k
⊞ Viewer · Updated Apr 22, 2024 · ▤ 209k · ↓2.11k · ♡ 107

casey-martin/multilingual-mathematical-autoformaliz…
⊞ Viewer · Updated Dec 3, 2023 · ▤ 666k · ↓275 · ♡ 3

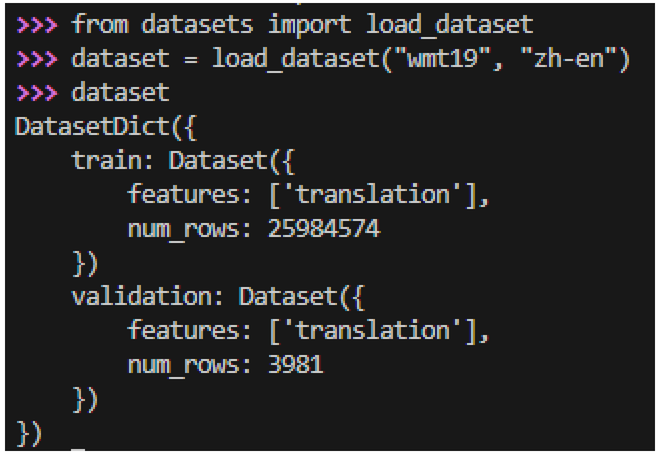
ReadingTimeMachine/rtm-sgt-ocr-v1
⊞ Viewer · Updated Jul 8 · ▤ 1.53M · ↓182 · ♡ 4

CohereLabs/aya_collection
⊞ Viewer · Updated Apr 15 · ▤ 514M · ↓7.3k · ♡ 227

ai4privacy/pii-masking-300k
⊞ Viewer · Updated Apr 22, 2024 · ▤ 225k · ↓2.4k · ♡ 69

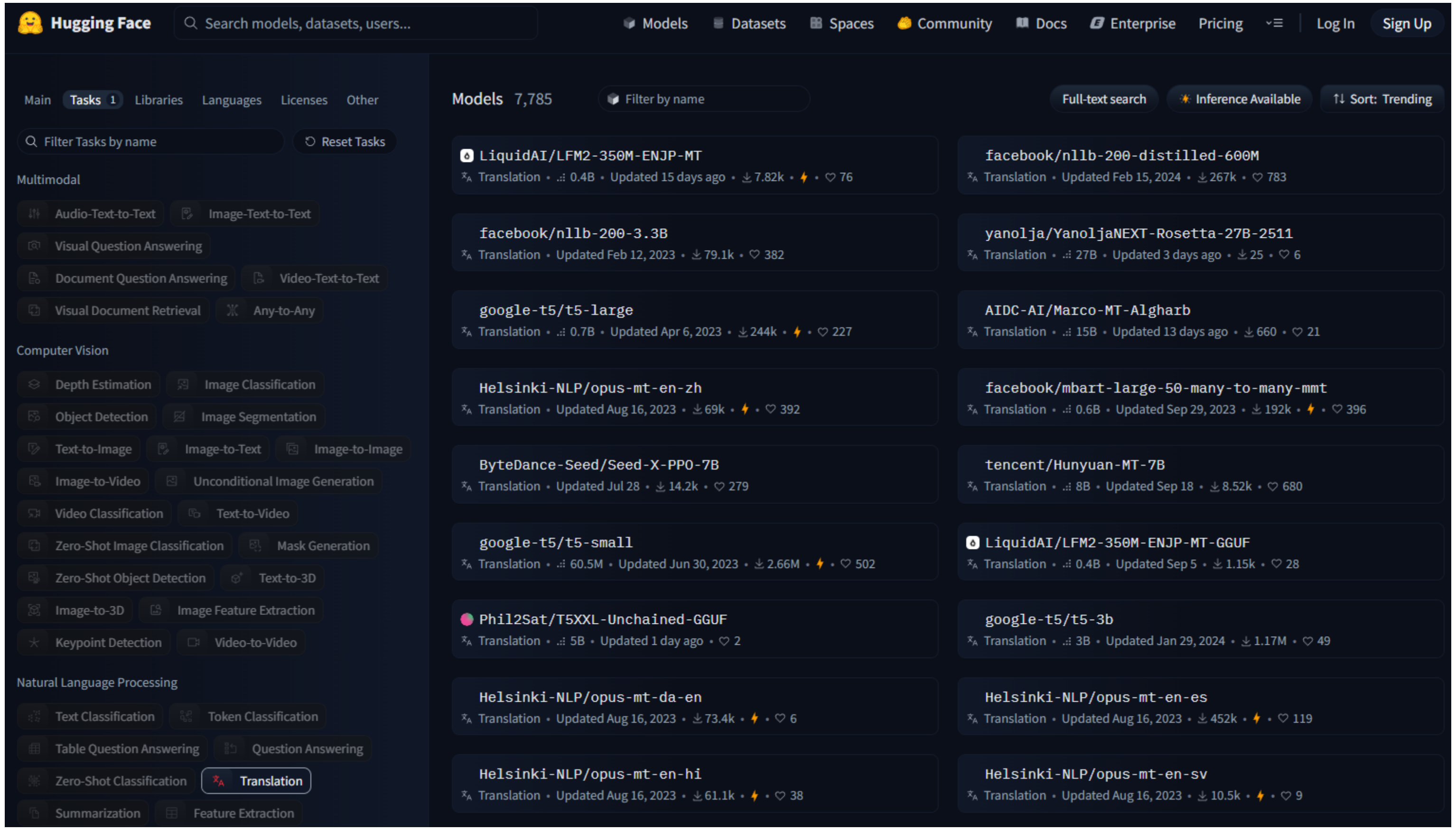
lmg-anon/vntl-leaderboard
⊞ Viewer · Updated Jan 3 · ▤ 87 · ↓351 · ♡ 39

Use one line to load datasets from HugginFace platform:



```
>>> from datasets import load_dataset
>>> dataset = load_dataset("wmt19", "zh-en")
>>> dataset
DatasetDict({
    train: Dataset({
        features: ['translation'],
        num_rows: 25984574
    })
    validation: Dataset({
        features: ['translation'],
        num_rows: 3981
    })
})
```

HuggingFace 🤗 Transformers:



LiquidAI/LFM2-350M-ENJP-MT
Translation • 0.4B • Updated 15 days ago • 7.82k • • 76

facebook/nllb-200-distilled-600M
Translation • Updated Feb 15, 2024 • 267k • 783

facebook/nllb-200-3.3B
Translation • Updated Feb 12, 2023 • 79.1k • 382

yanolja/YanoljaNEXT-Rosetta-27B-2511
Translation • 27B • Updated 3 days ago • 25 • 6

google-t5/t5-large
Translation • 0.7B • Updated Apr 6, 2023 • 244k • • 227

AIDC-AI/Marco-MT-Algharb
Translation • 15B • Updated 13 days ago • 660 • 21

Helsinki-NLP/opus-mt-en-zh
Translation • Updated Aug 16, 2023 • 69k • • 392

facebook/mbart-large-50-many-to-many-mmt
Translation • 0.6B • Updated Sep 29, 2023 • 192k • • 396

ByteDance-Seed/Seed-X-PPO-7B
Translation • Updated Jul 28 • 14.2k • 279

tencent/Hunyuan-MT-7B
Translation • 8B • Updated Sep 18 • 8.52k • 680

google-t5/t5-small
Translation • 60.5M • Updated Jun 30, 2023 • 2.66M • • 502

LiquidAI/LFM2-350M-ENJP-MT-GGUF
Translation • 0.4B • Updated Sep 5 • 1.15k • 28

Phil2Sat/T5XXL-Unchained-GGUF
Translation • 5B • Updated 1 day ago • 2

google-t5/t5-3b
Translation • 3B • Updated Jan 29, 2024 • 1.17M • 49

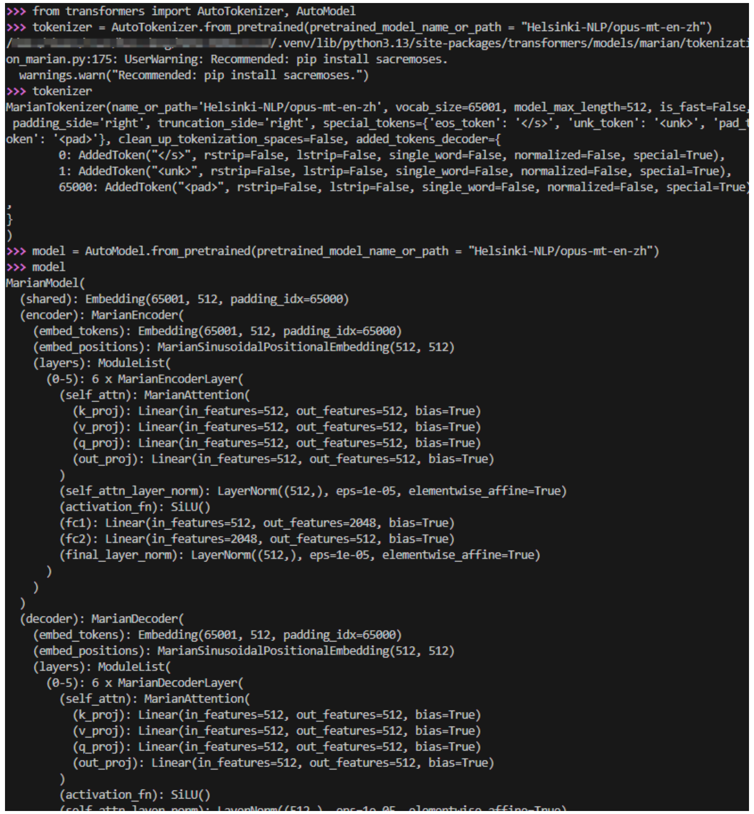
Helsinki-NLP/opus-mt-da-en
Translation • Updated Aug 16, 2023 • 73.4k • • 6

Helsinki-NLP/opus-mt-en-es
Translation • Updated Aug 16, 2023 • 452k • • 119

Helsinki-NLP/opus-mt-en-hi
Translation • Updated Aug 16, 2023 • 61.1k • • 38

Helsinki-NLP/opus-mt-en-sv
Translation • Updated Aug 16, 2023 • 10.5k • • 9

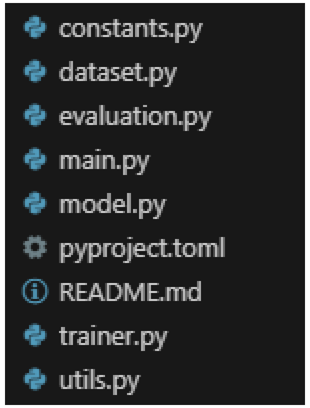
Use one line to load tokenizers and pretrained models from HugginFace platform:



```
>>> from transformers import AutoTokenizer, AutoModel
>>> tokenizer = AutoTokenizer.from_pretrained(pretrained_model_name_or_path = "Helsinki-NLP/opus-mt-en-zh")
/[redacted]/.venv/lib/python3.13/site-packages/transformers/models/marian/tokenizati
on_marian.py:175: UserWarning: Recommended: pip install sacremoses.
  warnings.warn("Recommended: pip install sacremoses.")
>>> tokenizer
MarianTokenizer(name_or_path='Helsinki-NLP/opus-mt-en-zh', vocab_size=65001, model_max_length=512, is_fast=False,
 padding_side='right', truncation_side='right', special_tokens={'eos_token': '</s>', 'unk_token': '<unk>', 'pad_t
oken': '<pad>'}, clean_up_tokenization_spaces=False, added_tokens_decoder={
        0: AddedToken("</s>", rstrip=False, lstrip=False, single_word=False, normalized=False, special=True),
        1: AddedToken("<unk>", rstrip=False, lstrip=False, single_word=False, normalized=False, special=True),
        65000: AddedToken("<pad>", rstrip=False, lstrip=False, single_word=False, normalized=False, special=True)
,
}
)
>>> model = AutoModel.from_pretrained(pretrained_model_name_or_path = "Helsinki-NLP/opus-mt-en-zh")
>>> model
MarianModel(
  (shared): Embedding(65001, 512, padding_idx=65000)
  (encoder): MarianEncoder(
    (embed_tokens): Embedding(65001, 512, padding_idx=65000)
    (embed_positions): MarianSinusoidalPositionalEmbedding(512, 512)
    (layers): ModuleList(
      (0-5): 6 x MarianEncoderLayer(
        (self_attn): MarianAttention(
          (k_proj): Linear(in_features=512, out_features=512, bias=True)
          (v_proj): Linear(in_features=512, out_features=512, bias=True)
          (q_proj): Linear(in_features=512, out_features=512, bias=True)
          (out_proj): Linear(in_features=512, out_features=512, bias=True)
        )
        (self_attn_layer_norm): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
        (activation_fn): SiLU()
        (fc1): Linear(in_features=512, out_features=2048, bias=True)
        (fc2): Linear(in_features=2048, out_features=512, bias=True)
        (final_layer_norm): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
      )
    )
  )
  (decoder): MarianDecoder(
    (embed_tokens): Embedding(65001, 512, padding_idx=65000)
    (embed_positions): MarianSinusoidalPositionalEmbedding(512, 512)
    (layers): ModuleList(
      (0-5): 6 x MarianDecoderLayer(
        (self_attn): MarianAttention(
          (k_proj): Linear(in_features=512, out_features=512, bias=True)
          (v_proj): Linear(in_features=512, out_features=512, bias=True)
          (q_proj): Linear(in_features=512, out_features=512, bias=True)
          (out_proj): Linear(in_features=512, out_features=512, bias=True)
        )
        (activation_fn): SiLU()
        (self_attn_layer_norm): LayerNorm((512,), eps=1e-05, elementwise_affine=True)
```

What we will do:

The workflow is organized into 5 files:


constants.py
dataset.py
evaluation.py
main.py
model.py
pyproject.toml
README.md
trainer.py
utils.py

- **Dataset Sourcing & Preparation (dataset.py)**
  Explore HF Datasets, load splits, perform filtering/mapping, and train/validation/test preparation

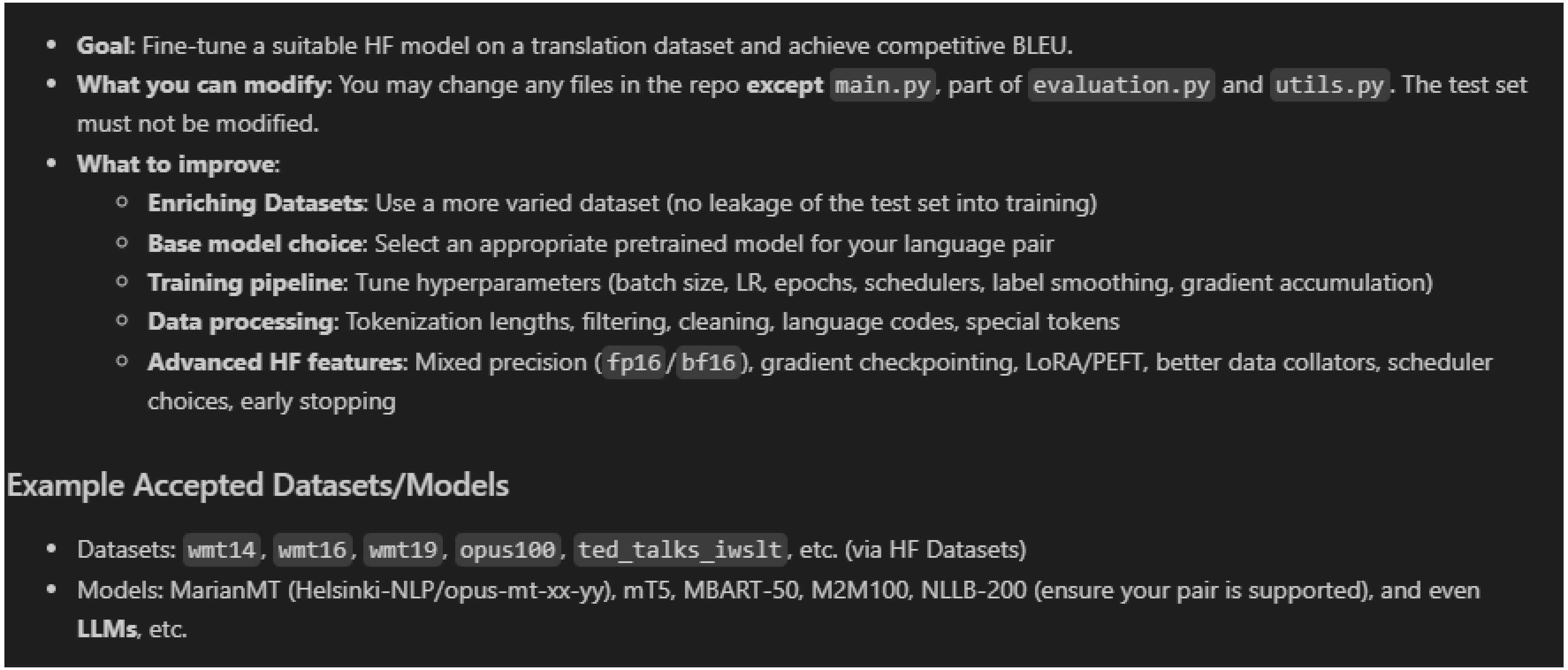
- **Model & Tokenizer Setup (model.py)**
  Select a suitable base model from the Hub and initialize tokenizer and config

- **Training Pipeline (trainer.py)**
  Configure TrainingArguments, data collators, metrics, logging, and checkpointing

- **Run & Monitor Training (main.py)**
  Orchestrate end-to-end training and validation, with periodic evaluation

- **Evaluation & Reporting (evaluation.py)**
  Compute BLEU on the held-out test set; save artifacts and summary

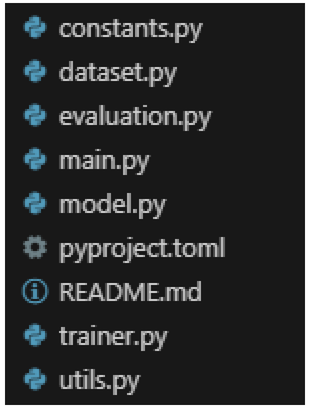
What we will do:



- **Goal**: Fine-tune a suitable HF model on a translation dataset and achieve competitive BLEU.
- **What you can modify**: You may change any files in the repo **except** `main.py`, part of `evaluation.py` and `utils.py`. The test set must not be modified.
- **What to improve**:
  - **Enriching Datasets**: Use a more varied dataset (no leakage of the test set into training)
  - **Base model choice**: Select an appropriate pretrained model for your language pair
  - **Training pipeline**: Tune hyperparameters (batch size, LR, epochs, schedulers, label smoothing, gradient accumulation)
  - **Data processing**: Tokenization lengths, filtering, cleaning, language codes, special tokens
  - **Advanced HF features**: Mixed precision (`fp16`/`bf16`), gradient checkpointing, LoRA/PEFT, better data collators, scheduler choices, early stopping

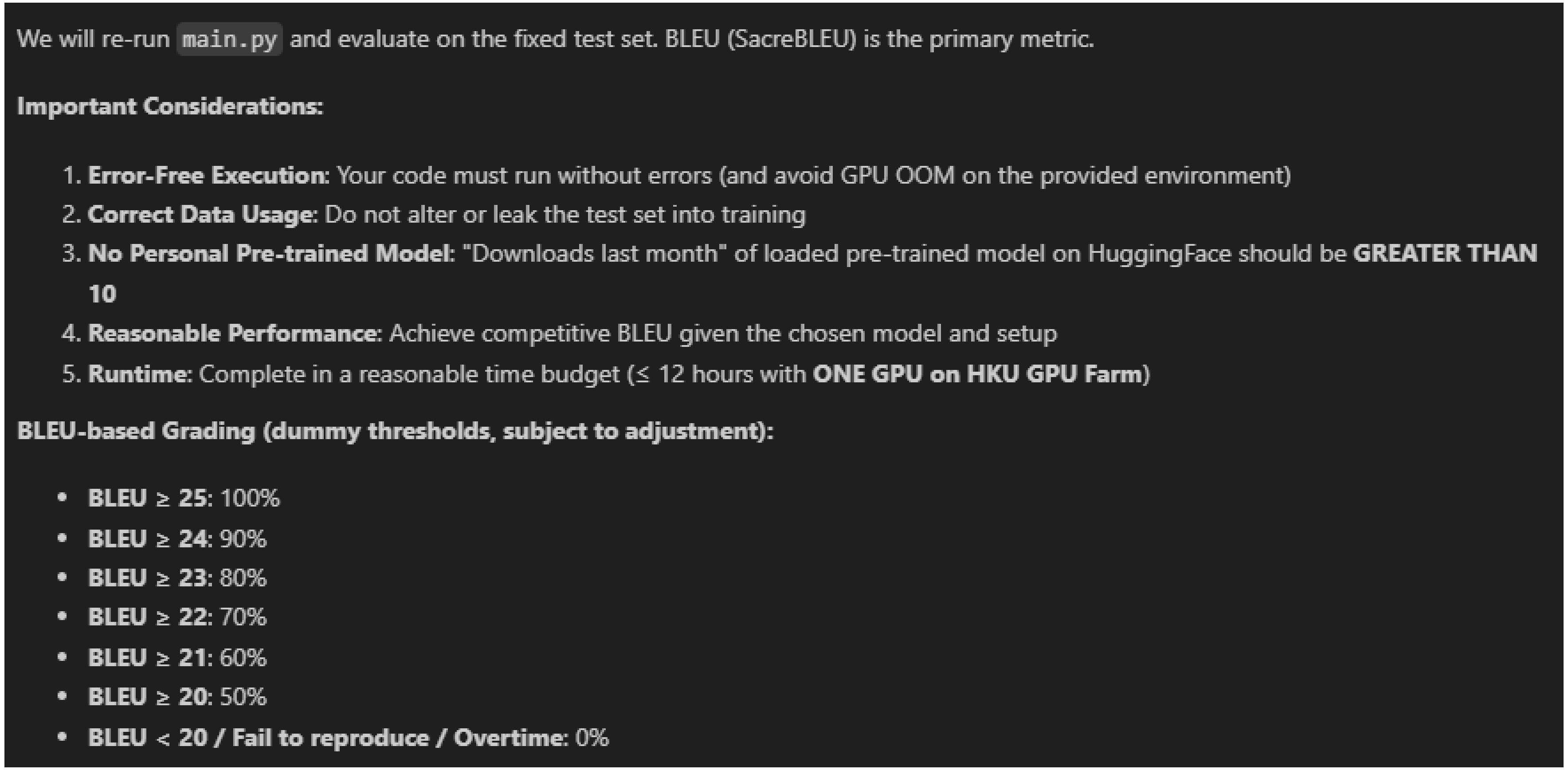
Example Accepted Datasets/Models

- Datasets: `wmt14`, `wmt16`, `wmt19`, `opus100`, `ted_talks_iwslt`, etc. (via HF Datasets)
- Models: MarianMT (Helsinki-NLP/opus-mt-xx-yy), mT5, MBART-50, M2M100, NLLB-200 (ensure your pair is supported), and even **LLMs**, etc.



constants.py
dataset.py
evaluation.py
main.py
model.py
pyproject.toml
README.md
trainer.py
utils.py

Grading:



We will re-run `main.py` and evaluate on the fixed test set. BLEU (SacreBLEU) is the primary metric.

Important Considerations:

1. **Error-Free Execution:** Your code must run without errors (and avoid GPU OOM on the provided environment)
2. **Correct Data Usage**: Do not alter or leak the test set into training
3. **No Personal Pre-trained Model**: "Downloads last month" of loaded pre-trained model on HuggingFace should be **GREATER THAN 10**
4. **Reasonable Performance**: Achieve competitive BLEU given the chosen model and setup
5. **Runtime**: Complete in a reasonable time budget (≤ 12 hours with **ONE GPU on HKU GPU Farm**)

BLEU-based Grading (dummy thresholds, subject to adjustment):

- **BLEU ≥ 25**: 100%
- **BLEU ≥ 24**: 90%
- **BLEU ≥ 23**: 80%
- **BLEU ≥ 22**: 70%
- **BLEU ≥ 21**: 60%
- **BLEU ≥ 20**: 50%
- **BLEU < 20 / Fail to reproduce / Overtime**: 0%

Important dates:

- Assignment II Release: Nov. 06 (Thursday)
- Submission Deadline: Nov. 30 (Sunday) (23:59 GMT+8)

Late submission policy :

- All submissions later than the deadline will NOT be accepted

Questions!

If any more questions, please contact kendwj@hku.hk
or schen59@hku.hk