# Use pre-trained BERT model to predict NER labels

see `CAVEAT` cell below

- This cell will download a large (400+MB) pre-trained model and likely take several minutes to do so.
- The Bert model will be saved to the local filesystem and not downloaded on subsequent invocations.

In [1]:
```
import time
ts = time.time()
```

In [10]:
```
%pip install datasets
%pip install evaluate
%pip install transformers
%pip install seqeval
```

```
Requirement already satisfied: datasets in /opt/conda/lib/python3.10/site-pa
ckages (2.16.1)
Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-pa
ckages (from datasets) (3.13.1)
Requirement already satisfied: numpy>=1.17 in /opt/conda/lib/python3.10/site
-packages (from datasets) (1.26.2)
Requirement already satisfied: pyarrow>=8.0.0 in /opt/conda/lib/python3.10/s
ite-packages (from datasets) (15.0.0)
Requirement already satisfied: pyarrow-hotfix in /opt/conda/lib/python3.10/s
ite-packages (from datasets) (0.6)
Requirement already satisfied: dill<0.3.8,>=0.3.0 in /opt/conda/lib/python3.
10/site-packages (from datasets) (0.3.7)
Requirement already satisfied: pandas in /opt/conda/lib/python3.10/site-pack
ages (from datasets) (2.2.0)
Requirement already satisfied: requests>=2.19.0 in /opt/conda/lib/python3.1
0/site-packages (from datasets) (2.31.0)
Requirement already satisfied: tqdm>=4.62.1 in /opt/conda/lib/python3.10/sit
e-packages (from datasets) (4.65.0)
Requirement already satisfied: xxhash in /opt/conda/lib/python3.10/site-pack
ages (from datasets) (3.4.1)
Requirement already satisfied: multiprocess in /opt/conda/lib/python3.10/sit
e-packages (from datasets) (0.70.15)
Requirement already satisfied: fsspec<=2023.10.0,>=2023.1.0 in /opt/conda/li
b/python3.10/site-packages (from fsspec[http]<=2023.10.0,>=2023.1.0->dataset
s) (2023.10.0)
Requirement already satisfied: aiohttp in /opt/conda/lib/python3.10/site-pac
kages (from datasets) (3.9.2)
Requirement already satisfied: huggingface-hub>=0.19.4 in /opt/conda/lib/pyt
hon3.10/site-packages (from datasets) (0.20.3)
Requirement already satisfied: packaging in /opt/conda/lib/python3.10/site-p
ackages (from datasets) (23.1)
Requirement already satisfied: pyyaml>=5.1 in /opt/conda/lib/python3.10/site
-packages (from datasets) (6.0.1)
Requirement already satisfied: aiosignal>=1.1.2 in /opt/conda/lib/python3.1
0/site-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /opt/conda/lib/python3.10/si
te-packages (from aiohttp->datasets) (23.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in /opt/conda/lib/python3.1
0/site-packages (from aiohttp->datasets) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /opt/conda/lib/python
3.10/site-packages (from aiohttp->datasets) (6.0.4)
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/conda/lib/python3.10/s
ite-packages (from aiohttp->datasets) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /opt/conda/lib/pyt
hon3.10/site-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /opt/conda/lib/
python3.10/site-packages (from huggingface-hub>=0.19.4->datasets) (4.7.1)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/py
thon3.10/site-packages (from requests>=2.19.0->datasets) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/sit
e-packages (from requests>=2.19.0->datasets) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/lib/python3.
10/site-packages (from requests>=2.19.0->datasets) (1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.
10/site-packages (from requests>=2.19.0->datasets) (2023.11.17)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/conda/lib/pyth
```

```
on3.10/site-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/sit
e-packages (from pandas->datasets) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.7 in /opt/conda/lib/python3.10/s
ite-packages (from pandas->datasets) (2023.4)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.10/site-pa
ckages (from python-dateutil>=2.8.2->pandas->datasets) (1.16.0)
WARNING: Running pip as the 'root' user can result in broken permissions and
conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: https://pip.pypa.io/warnings/venv
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: evaluate in /opt/conda/lib/python3.10/site-pa
ckages (0.4.1)
Requirement already satisfied: datasets>=2.0.0 in /opt/conda/lib/python3.10/
site-packages (from evaluate) (2.16.1)
Requirement already satisfied: numpy>=1.17 in /opt/conda/lib/python3.10/site
-packages (from evaluate) (1.26.2)
Requirement already satisfied: dill in /opt/conda/lib/python3.10/site-packag
es (from evaluate) (0.3.7)
Requirement already satisfied: pandas in /opt/conda/lib/python3.10/site-pack
ages (from evaluate) (2.2.0)
Requirement already satisfied: requests>=2.19.0 in /opt/conda/lib/python3.1
0/site-packages (from evaluate) (2.31.0)
Requirement already satisfied: tqdm>=4.62.1 in /opt/conda/lib/python3.10/sit
e-packages (from evaluate) (4.65.0)
Requirement already satisfied: xxhash in /opt/conda/lib/python3.10/site-pack
ages (from evaluate) (3.4.1)
Requirement already satisfied: multiprocess in /opt/conda/lib/python3.10/sit
e-packages (from evaluate) (0.70.15)
Requirement already satisfied: fsspec>=2021.05.0 in /opt/conda/lib/python3.1
0/site-packages (from fsspec[http]>=2021.05.0->evaluate) (2023.10.0)
Requirement already satisfied: huggingface-hub>=0.7.0 in /opt/conda/lib/pyth
on3.10/site-packages (from evaluate) (0.20.3)
Requirement already satisfied: packaging in /opt/conda/lib/python3.10/site-p
ackages (from evaluate) (23.1)
Requirement already satisfied: responses<0.19 in /opt/conda/lib/python3.10/s
ite-packages (from evaluate) (0.18.0)
Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-pa
ckages (from datasets>=2.0.0->evaluate) (3.13.1)
Requirement already satisfied: pyarrow>=8.0.0 in /opt/conda/lib/python3.10/s
ite-packages (from datasets>=2.0.0->evaluate) (15.0.0)
Requirement already satisfied: pyarrow-hotfix in /opt/conda/lib/python3.10/s
ite-packages (from datasets>=2.0.0->evaluate) (0.6)
Requirement already satisfied: aiohttp in /opt/conda/lib/python3.10/site-pac
kages (from datasets>=2.0.0->evaluate) (3.9.2)
Requirement already satisfied: pyyaml>=5.1 in /opt/conda/lib/python3.10/site
-packages (from datasets>=2.0.0->evaluate) (6.0.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /opt/conda/lib/
python3.10/site-packages (from huggingface-hub>=0.7.0->evaluate) (4.7.1)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/py
thon3.10/site-packages (from requests>=2.19.0->evaluate) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/sit
e-packages (from requests>=2.19.0->evaluate) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/lib/python3.
10/site-packages (from requests>=2.19.0->evaluate) (1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.
```

```
10/site-packages (from requests>=2.19.0->evaluate) (2023.11.17)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/conda/lib/pyth
on3.10/site-packages (from pandas->evaluate) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/sit
e-packages (from pandas->evaluate) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.7 in /opt/conda/lib/python3.10/s
ite-packages (from pandas->evaluate) (2023.4)
Requirement already satisfied: aiosignal>=1.1.2 in /opt/conda/lib/python3.1
0/site-packages (from aiohttp->datasets>=2.0.0->evaluate) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /opt/conda/lib/python3.10/si
te-packages (from aiohttp->datasets>=2.0.0->evaluate) (23.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in /opt/conda/lib/python3.1
0/site-packages (from aiohttp->datasets>=2.0.0->evaluate) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /opt/conda/lib/python
3.10/site-packages (from aiohttp->datasets>=2.0.0->evaluate) (6.0.4)
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/conda/lib/python3.10/s
ite-packages (from aiohttp->datasets>=2.0.0->evaluate) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /opt/conda/lib/pyt
hon3.10/site-packages (from aiohttp->datasets>=2.0.0->evaluate) (4.0.3)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.10/site-pa
ckages (from python-dateutil>=2.8.2->pandas->evaluate) (1.16.0)
WARNING: Running pip as the 'root' user can result in broken permissions and
conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: https://pip.pypa.io/warnings/venv
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: transformers in /opt/conda/lib/python3.10/sit
e-packages (4.37.1)
Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-pa
ckages (from transformers) (3.13.1)
Requirement already satisfied: huggingface-hub<1.0,>=0.19.3 in /opt/conda/li
b/python3.10/site-packages (from transformers) (0.20.3)
Requirement already satisfied: numpy>=1.17 in /opt/conda/lib/python3.10/site
-packages (from transformers) (1.26.2)
Requirement already satisfied: packaging>=20.0 in /opt/conda/lib/python3.10/
site-packages (from transformers) (23.1)
Requirement already satisfied: pyyaml>=5.1 in /opt/conda/lib/python3.10/site
-packages (from transformers) (6.0.1)
Requirement already satisfied: regex!=2019.12.17 in /opt/conda/lib/python3.1
0/site-packages (from transformers) (2023.12.25)
Requirement already satisfied: requests in /opt/conda/lib/python3.10/site-pa
ckages (from transformers) (2.31.0)
Requirement already satisfied: tokenizers<0.19,>=0.14 in /opt/conda/lib/pyth
on3.10/site-packages (from transformers) (0.15.1)
Requirement already satisfied: safetensors>=0.3.1 in /opt/conda/lib/python3.
10/site-packages (from transformers) (0.4.2)
Requirement already satisfied: tqdm>=4.27 in /opt/conda/lib/python3.10/site-
packages (from transformers) (4.65.0)
Requirement already satisfied: fsspec>=2023.5.0 in /opt/conda/lib/python3.1
0/site-packages (from huggingface-hub<1.0,>=0.19.3->transformers) (2023.10.
0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /opt/conda/lib/
python3.10/site-packages (from huggingface-hub<1.0,>=0.19.3->transformers)
(4.7.1)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/py
thon3.10/site-packages (from requests->transformers) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/sit
```

```
e-packages (from requests->transformers) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in /opt/conda/lib/python3.
10/site-packages (from requests->transformers) (1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.
10/site-packages (from requests->transformers) (2023.11.17)
WARNING: Running pip as the 'root' user can result in broken permissions and
conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: https://pip.pypa.io/warnings/venv
Note: you may need to restart the kernel to use updated packages.
Collecting seqeval
  Downloading seqeval-1.2.2.tar.gz (43 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.6/43.6 kB 1.0 MB/s eta 0:0
0:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: numpy>=1.14.0 in /opt/conda/lib/python3.10/si
te-packages (from seqeval) (1.26.2)
Requirement already satisfied: scikit-learn>=0.21.3 in /opt/conda/lib/python
3.10/site-packages (from seqeval) (1.4.0)
Requirement already satisfied: scipy>=1.6.0 in /opt/conda/lib/python3.10/sit
e-packages (from scikit-learn>=0.21.3->seqeval) (1.12.0)
Requirement already satisfied: joblib>=1.2.0 in /opt/conda/lib/python3.10/si
te-packages (from scikit-learn>=0.21.3->seqeval) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /opt/conda/lib/python
3.10/site-packages (from scikit-learn>=0.21.3->seqeval) (3.2.0)
Building wheels for collected packages: seqeval
  Building wheel for seqeval (setup.py) ... done
  Created wheel for seqeval: filename=seqeval-1.2.2-py3-none-any.whl size=16
162 sha256=b1d42997f4ef42791b067d0a163f045345af77d449348bb3baad40124f3c0d47
  Stored in directory: /root/.cache/pip/wheels/1a/67/4a/ad4082dd7dfc30f2abfe
4d80a2ed5926a506eb8a972b4767fa
Successfully built seqeval
Installing collected packages: seqeval
Successfully installed seqeval-1.2.2
WARNING: Running pip as the 'root' user can result in broken permissions and
conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: https://pip.pypa.io/warnings/venv
Note: you may need to restart the kernel to use updated packages.
```

In [3]:
```
from utils import *
```

In [4]:
```
# import needed dependences
from transformers import pipeline
from transformers import AutoTokenizer
from transformers import AutoModelForTokenClassification
from transformers.pipelines.token_classification import TokenClassificationP
```

In [5]:
```
# CAVEAT: Long-running cell on initial invocation
# load pre-trained BERT base cased model
tokenizer = AutoTokenizer.from_pretrained("dslim/bert-base-NER")
bert_model = AutoModelForTokenClassification.from_pretrained("dslim/bert-bas
bert_nlp: TokenClassificationPipeline = pipeline("ner", model=bert_model, to
```

```
tokenizer_config.json:   0%|          | 0.00/59.0 [00:00<?, ?B/s]
config.json:   0%|          | 0.00/829 [00:00<?, ?B/s]
vocab.txt:   0%|          | 0.00/213k [00:00<?, ?B/s]
added_tokens.json:   0%|          | 0.00/2.00 [00:00<?, ?B/s]
```

```
special_tokens_map.json:   0%|          | 0.00/112 [00:00<?, ?B/s]
model.safetensors:   0%|          | 0.00/433M [00:00<?, ?B/s]
```

```
Some weights of the model checkpoint at dslim/bert-base-NER were not used wh
en initializing BertForTokenClassification: ['bert.pooler.dense.bias', 'ber
t.pooler.dense.weight']
- This IS expected if you are initializing BertForTokenClassification from t
he checkpoint of a model trained on another task or with another architectur
e (e.g. initializing a BertForSequenceClassification model from a BertForPre
Training model).
- This IS NOT expected if you are initializing BertForTokenClassification fr
om the checkpoint of a model that you expect to be exactly identical (initia
lizing a BertForSequenceClassification model from a BertForSequenceClassific
ation model).
```

In [6]:
```python
def annotate_bert(result: int) -> str:
    """ Helper function to tranlate NER int to class label """
    if len(result) == 0:
        return 'O'
    return result[0]['entity']

def predict_ner_bert(bert_nlp: TokenClassificationPipeline, labeled_dataset:
    """ Run inference on the tokens using trained BERT model """
    references: list[list[str]] = []
    predictions: list[list[str]] = []

    for row in tqdm(labeled_dataset, desc=str(len(labeled_dataset))):
        # add ground truth labels to references
        references.append([tag_names[id] for id in row['ner_tags']])
        # recognize named entity in a test tokens
        ner_results = bert_nlp(row['tokens'])
        # translate numerical index to NER class label
        predicted_tags = [annotate_bert(x) for x in ner_results]
        predictions.append(predicted_tags)
    return references, predictions
```

In [7]:
```python
# Generate a list of ground truth NER labels and predictions
references, bert_predictions = predict_ner_bert(bert_nlp, test)
```

```
3453: 100%|██████████| 3453/3453 [1:04:22<00:00,  1.12s/it]
```

In [8]:
```python
# Save NER results to disk
bert_path = os.path.join(interim_dir, "ner_results_bert.json")
save_ner_results(bert_path, references, bert_predictions)

# Load persisted NER results
# references1, predictions1 = load_ner_results(bert_path)
```

```
Saving NER results to ../data/interim/ner_results_bert.json
```

In [11]:
```python
# Evaluate results
results = evaluate_results(references, bert_predictions)

bert_evaluation_path = os.path.join(interim_dir, "evaluation_results_bert.js
save_evaluation_results(bert_evaluation_path, results)
```

```
Saving evaluation results to ../data/interim/evaluation_results_bert.json
```

In [13]:
```
te = time.time()
duration = te-ts
duration = float(f"{duration:.2f}")
print(f"Total running time: {duration} sec ")
```

```
Total running time: 3947.72 sec
```

In [ ]:
```

```