Stanford NER tagger w. NLTK

In [1]:
```
%pip install datasets
%pip install evaluate
%pip install nltk
```

```
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: nltk in /opt/conda/lib/python3.10/site-packag
es (3.8.1)
Requirement already satisfied: click in /opt/conda/lib/python3.10/site-packa
ges (from nltk) (8.1.7)
Requirement already satisfied: joblib in /opt/conda/lib/python3.10/site-pack
ages (from nltk) (1.3.2)
Requirement already satisfied: regex>=2021.8.3 in /opt/conda/lib/python3.10/
site-packages (from nltk) (2023.12.25)
Requirement already satisfied: tqdm in /opt/conda/lib/python3.10/site-packag
es (from nltk) (4.65.0)
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.
```

In [10]:
```
%pip install seqeval
```

```
Collecting seqeval
  Downloading seqeval-1.2.2.tar.gz (43 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.6/43.6 kB 3.8 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)
Collecting scikit-learn>=0.21.3 (from seqeval)
  Downloading scikit_learn-1.4.0-1-cp310-cp310-manylinux_2_17_x86_64.manylin
ux2014_x86_64.whl.metadata (11 kB)
Collecting scipy>=1.6.0 (from scikit-learn>=0.21.3->seqeval)
  Downloading scipy-1.12.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x
86_64.whl.metadata (60 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60.4/60.4 kB 8.6 MB/s eta 0:0
0:00
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)
Collecting threadpoolctl>=2.0.0 (from scikit-learn>=0.21.3->seqeval)
  Downloading threadpoolctl-3.2.0-py3-none-any.whl.metadata (10.0 kB)
Downloading scikit_learn-1.4.0-1-cp310-cp310-manylinux_2_17_x86_64.manylinux
2014_x86_64.whl (12.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.1/12.1 MB 49.8 MB/s eta 0:00:
00a 0:00:01
Downloading scipy-1.12.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86
_64.whl (38.4 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 38.4/38.4 MB 129.3 MB/s eta 0:0
0:0000:0100:01
Downloading threadpoolctl-3.2.0-py3-none-any.whl (15 kB)
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=3a72830b590d3879461b45bf4fe2e6b7a65678bba07eec3cdb581e1c7263841d
  Stored in directory: /root/.cache/pip/wheels/1a/67/4a/ad4082dd7dfc30f2abfe
4d80a2ed5926a506eb8a972b4767fa
Successfully built seqeval
Installing collected packages: threadpoolctl, scipy, scikit-learn, seqeval
Successfully installed scikit-learn-1.4.0 scipy-1.12.0 seqeval-1.2.2 threadp
oolctl-3.2.0
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.
```

In [2]:
```
# %%bash
# wget http://nlp.stanford.edu/software/stanford-ner-2015-04-20.zip
# unzip stanford-ner-2015-04-20.zip
```

In [3]:
```
from utils import *
from pathlib import Path
import time
ts = time.time()
```

In [4]:
```
from nltk.tag.stanford import StanfordNERTagger

CWD = Path().resolve()
```

```
stanford_dir = os.path.join(CWD, "stanford-ner-2015-04-20")
jar = os.path.join(stanford_dir, "stanford-ner-3.5.2.jar")
stanford_model_path = os.path.join(stanford_dir, "classifiers")
```

In [5]:
```
stanford_model = os.path.join(stanford_model_path, "english.conll.4class.dis
stanford_tagger: StanfordNERTagger = StanfordNERTagger(stanford_model, jar,
```

In [6]:
```
def annotate_stanford(result: str) -> str:
    """ Helper function to tranlate NER int to class label """
    match result:
        case "LOCATION":
            return "LOC"
        case "PERSON":
            return "PER"
        case "ORGANIZATION":
            return "ORG"
        case "MISC":
            return "MISC"
        case "O":
            return "O"
        case _:
            return "X"

def predict_ner_stanford(stanford_tagger: StanfordNERTagger, labeled_dataset
    """ Run inference on the tokens using trained BERT model """
    references: list[list[str]] = []
    st_predictions: list[list[str]] = []

    for row in tqdm(labeled_dataset, desc=str(len(labeled_dataset))):
        # add ground truth labels to references
        references.append([re.sub("^[BI]-", "", tag_names[id]) for id in row
        # recognize named entity in a test tokens
        ner_results = stanford_tagger.tag(row['tokens'])
        # translate numerical index to NER class label
        predicted_tags = [annotate_stanford(y) for x, y in ner_results]
        st_predictions.append(predicted_tags)
    return references, st_predictions
```

In [7]:
```
# Run NER inference using Stanford NER tagger
references, st_predictions = predict_ner_stanford(stanford_tagger, test)
```

```
3453: 100%|██████████| 3453/3453 [1:01:42<00:00,  1.07s/it]
```

In [8]:
```
# Save NER results to disk
stanford_results_path = os.path.join(interim_dir, "ner_results_stanford.jsor
save_ner_results(stanford_results_path, references, st_predictions)

# Load persisted NER results
# references, st_predictions = load_ner_results(stanford_results_path)
```

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

In [11]:
```
results = evaluate_results(references, st_predictions)
results
```

Out[11]:

```
{'ER': {'precision': 0.9509360877985797,
  'recall': 0.9109461966604824,
  'f1': 0.930511686670878,
  'number': 1617},
 'ISC': {'precision': 0.8191027496382055,
  'recall': 0.8167388167388168,
  'f1': 0.8179190751445087,
  'number': 693},
 'OC': {'precision': 0.9049881235154394,
  'recall': 0.9169675090252708,
  'f1': 0.9109384339509863,
  'number': 1662},
 'overall_precision': 0.9080020387359837,
 'overall_recall': 0.8970292044310171,
 'overall_f1': 0.902482269503546,
 'overall_accuracy': 0.9758587272531496}
```


In [12]:
```
stanford_evaluation_path = os.path.join(interim_dir, "evaluation_results_sta
save_evaluation_results(stanford_evaluation_path, results)
```

```
Saving evaluation results to ../data/interim/evaluation_results_stanford.jso
n
```

In [14]:
```
te = time.time()
duration = te-ts
duration = float(f"{duration:.2f}")
print(f"Total running time: {duration} sec ")
```

```
Total running time: 4013.37 sec
```

In [ ]:
```

```