final5-8

May 4, 2025

[85]:

```
!pip install -U spacy
!python -m spacy download en_core_web_sm      #       ; en_core_web_trf -␣
↪Transformer-
```

```
Requirement already satisfied: spacy in /usr/local/lib/python3.11/dist-packages
(3.8.5)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in
/usr/local/lib/python3.11/dist-packages (from spacy) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in
/usr/local/lib/python3.11/dist-packages (from spacy) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/usr/local/lib/python3.11/dist-packages (from spacy) (1.0.12)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
/usr/local/lib/python3.11/dist-packages (from spacy) (2.0.11)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/usr/local/lib/python3.11/dist-packages (from spacy) (3.0.9)
Requirement already satisfied: thinc<8.4.0,>=8.3.4 in
/usr/local/lib/python3.11/dist-packages (from spacy) (8.3.6)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in
/usr/local/lib/python3.11/dist-packages (from spacy) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in
/usr/local/lib/python3.11/dist-packages (from spacy) (2.5.1)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in
/usr/local/lib/python3.11/dist-packages (from spacy) (2.0.10)
Requirement already satisfied: weasel<0.5.0,>=0.1.0 in
/usr/local/lib/python3.11/dist-packages (from spacy) (0.4.1)
Requirement already satisfied: typer<1.0.0,>=0.3.0 in
/usr/local/lib/python3.11/dist-packages (from spacy) (0.15.3)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in
/usr/local/lib/python3.11/dist-packages (from spacy) (4.67.1)
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.11/dist-
packages (from spacy) (2.0.2)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in
/usr/local/lib/python3.11/dist-packages (from spacy) (2.32.3)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in
/usr/local/lib/python3.11/dist-packages (from spacy) (2.11.3)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages
(from spacy) (3.1.6)
```

```
Requirement already satisfied: setuptools in /usr/local/lib/python3.11/dist-
packages (from spacy) (75.2.0)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.11/dist-packages (from spacy) (25.0)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in
/usr/local/lib/python3.11/dist-packages (from spacy) (3.5.0)
Requirement already satisfied: language-data>=1.2 in
/usr/local/lib/python3.11/dist-packages (from langcodes<4.0.0,>=3.2.0->spacy)
(1.3.0)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.11/dist-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.1 in
/usr/local/lib/python3.11/dist-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (2.33.1)
Requirement already satisfied: typing-extensions>=4.12.2 in
/usr/local/lib/python3.11/dist-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (4.13.2)
Requirement already satisfied: typing-inspection>=0.4.0 in
/usr/local/lib/python3.11/dist-packages (from
pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4->spacy) (0.4.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.13.0->spacy)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests<3.0.0,>=2.13.0->spacy) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.13.0->spacy)
(2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.13.0->spacy)
(2025.4.26)
Requirement already satisfied: blis<1.4.0,>=1.3.0 in
/usr/local/lib/python3.11/dist-packages (from thinc<8.4.0,>=8.3.4->spacy)
(1.3.0)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in
/usr/local/lib/python3.11/dist-packages (from thinc<8.4.0,>=8.3.4->spacy)
(0.1.5)
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.11/dist-
packages (from typer<1.0.0,>=0.3.0->spacy) (8.1.8)
Requirement already satisfied: shellingham>=1.3.0 in
/usr/local/lib/python3.11/dist-packages (from typer<1.0.0,>=0.3.0->spacy)
(1.5.4)
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.11/dist-
packages (from typer<1.0.0,>=0.3.0->spacy) (14.0.0)
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in
/usr/local/lib/python3.11/dist-packages (from weasel<0.5.0,>=0.1.0->spacy)
(0.21.0)
```

```
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in
/usr/local/lib/python3.11/dist-packages (from weasel<0.5.0,>=0.1.0->spacy)
(7.1.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->spacy) (3.0.2)
Requirement already satisfied: marisa-trie>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from language-
data>=1.2->langcodes<4.0.0,>=3.2.0->spacy) (1.2.1)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.11/dist-packages (from
rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/usr/local/lib/python3.11/dist-packages (from
rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy) (2.19.1)
Requirement already satisfied: wrapt in /usr/local/lib/python3.11/dist-packages
(from smart-open<8.0.0,>=5.2.1->weasel<0.5.0,>=0.1.0->spacy) (1.17.2)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-
packages (from markdown-it-py>=2.2.0->rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy)
(0.1.2)
Collecting en-core-web-sm==3.8.0
  Downloading https://github.com/explosion/spacy-
models/releases/download/en_core_web_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-
any.whl (12.8 MB)
                           12.8/12.8 MB
140.6 MB/s eta 0:00:00
 Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
 Restart to reload dependencies
If you are in a Jupyter or Colab notebook, you may need to restart Python in
order to load all the package's dependencies. You can do this by selecting the
'Restart kernel' or 'Restart runtime' option.
```

[86]:

```python
import spacy, json, pathlib
from spacy.scorer import Scorer
nlp = spacy.load("en_core_web_sm")

text = "Apple is looking at buying U.K. startup Graphcore for $1 billion."
doc  = nlp(text)

print("-- NER")
for ent in doc.ents:
    print(ent.text, ent.label_)

print("-- POS")
for token in doc:
    print(token.text, token.pos_, token.tag_)
```

```
-- NER
```

```
Apple ORG
U.K. GPE
Graphcore PERSON
$1 billion MONEY
-- POS
Apple PROPN NNP
is AUX VBZ
looking VERB VBG
at ADP IN
buying VERB VBG
U.K. PROPN NNP
startup VERB VBD
Graphcore PROPN NNP
for ADP IN
$ SYM $
1 NUM CD
billion NUM CD
. PUNCT .
```

[90]:

```python
import spacy, datasets
from spacy.tokens import Doc
from spacy.training import Example
from spacy.scorer import Scorer

# 1  load model and data
nlp   = spacy.load("en_core_web_sm")          # or _trf for better scores
conll = datasets.load_dataset("conll2003", split="validation")
id2label = conll.features["ner_tags"].feature.names

# 2  build gold-standard Docs with correct tokenisation
examples = []
for words, tags in zip(conll["tokens"], conll["ner_tags"]):
    # build a Doc with EXACT SAME WORD SEGMENTATION
    doc_gold = Doc(nlp.vocab, words=words)

    # convert BIO tags → spaCy entity spans
    spans = []
    start = None
    ent_type = None
    for i, t in enumerate(tags):
        tag = id2label[t]
        if tag == "O":
            if start is not None:
                spans.append((start, i, ent_type))
                start, ent_type = None, None
        elif tag.startswith("B-"):
            if start is not None:
                spans.append((start, i, ent_type))
            start, ent_type = i, tag[2:]
        elif tag.startswith("I-") and ent_type == tag[2:]:
            continue
        else:  # mismatched I-tag
            start, ent_type = None, None
    if start is not None:
        spans.append((start, len(tags), ent_type))

    doc_gold.ents = [doc_gold.char_span(doc_gold[i:j].start_char,
                                        doc_gold[i:j].end_char,
                                        label=l) for i, j, l in spans]

    doc_gold.ents = [e for e in doc_gold.ents if e is not None]

    doc_pred = nlp(doc_gold.text)
    examples.append(Example(doc_pred, doc_gold))

from spacy.scorer import Scorer
scorer   = Scorer()
results  = scorer.score(examples)

print("precision:", results["ents_p"])
print("recall   :", results["ents_r"])
print("f1       :", results["ents_f"])
```

```
precision: 0.05030812324929972
recall   : 0.07556378323796702
f1       : 0.060402233133786246
```

[3]:

```
!pip install datasets
!pip install evaluate
```

```
Collecting datasets
  Downloading datasets-3.5.1-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.18.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from datasets) (2.0.2)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (20.0.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(from datasets) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in
/usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.11/dist-
```

```
packages (from datasets) (4.67.1)
Collecting xxhash (from datasets)
  Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(12 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2 kB)
Collecting fsspec<=2025.3.0,>=2023.1.0 (from
fsspec[http]<=2025.3.0,>=2023.1.0->datasets)
  Downloading fsspec-2025.3.0-py3-none-any.whl.metadata (11 kB)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.11.15)
Requirement already satisfied: huggingface-hub>=0.24.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (0.30.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from datasets) (25.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from datasets) (6.0.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.6.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.4.3)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.20.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.24.0->datasets)
(4.13.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests>=2.32.2->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2025.4.26)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.9.0.post0)
```

```
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->datasets) (1.17.0)
Downloading datasets-3.5.1-py3-none-any.whl (491 kB)
                         491.4/491.4 kB
17.7 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                         116.3/116.3 kB
11.2 MB/s eta 0:00:00
Downloading fsspec-2025.3.0-py3-none-any.whl (193 kB)
                         193.6/193.6 kB
18.5 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py311-none-any.whl (143 kB)
                         143.5/143.5 kB
15.5 MB/s eta 0:00:00
Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
                         194.8/194.8 kB
20.0 MB/s eta 0:00:00
Installing collected packages: xxhash, fsspec, dill, multiprocess,
datasets
  Attempting uninstall: fsspec
    Found existing installation: fsspec 2025.3.2
    Uninstalling fsspec-2025.3.2:
      Successfully uninstalled fsspec-2025.3.2
Successfully installed datasets-3.5.1 dill-0.3.8 fsspec-2025.3.0
multiprocess-0.70.16 xxhash-3.5.0
Collecting evaluate
  Downloading evaluate-0.4.3-py3-none-any.whl.metadata (9.2 kB)
Requirement already satisfied: datasets>=2.0.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (3.5.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from evaluate) (2.0.2)
Requirement already satisfied: dill in /usr/local/lib/python3.11/dist-packages
(from evaluate) (0.3.8)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(from evaluate) (2.2.2)
Requirement already satisfied: requests>=2.19.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (2.32.3)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.11/dist-
packages (from evaluate) (4.67.1)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages
(from evaluate) (3.5.0)
Requirement already satisfied: multiprocess in /usr/local/lib/python3.11/dist-
packages (from evaluate) (0.70.16)
```

```
Requirement already satisfied: fsspec>=2021.05.0 in
/usr/local/lib/python3.11/dist-packages (from fsspec[http]>=2021.05.0->evaluate)
(2025.3.0)
Requirement already satisfied: huggingface-hub>=0.7.0 in
/usr/local/lib/python3.11/dist-packages (from evaluate) (0.30.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from evaluate) (25.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (3.18.0)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets>=2.0.0->evaluate)
(20.0.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (3.11.15)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from datasets>=2.0.0->evaluate) (6.0.2)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.7.0->evaluate)
(4.13.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests>=2.19.0->evaluate) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
(2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->evaluate)
(2025.4.26)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->evaluate) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->evaluate) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->evaluate) (2025.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->datasets>=2.0.0->evaluate) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.6.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
```

```
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (6.4.3)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from
aiohttp->datasets>=2.0.0->evaluate) (1.20.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->evaluate) (1.17.0)
Downloading evaluate-0.4.3-py3-none-any.whl (84 kB)
                         84.0/84.0 kB
5.7 MB/s eta 0:00:00
Installing collected packages: evaluate
Successfully installed evaluate-0.4.3
```

[4]:

```python
from datasets import load_dataset
conll = load_dataset("conll2003")
label_list = conll["train"].features["ner_tags"].feature.names
num_labels = len(label_list)
```

```
/usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94:
UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab
(https://huggingface.co/settings/tokens), set it as secret in your Google Colab
and restart your session.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access
public models or datasets.
  warnings.warn(

README.md:   0%|          | 0.00/12.3k [00:00<?, ?B/s]

conll2003.py:   0%|          | 0.00/9.57k [00:00<?, ?B/s]

The repository for conll2003 contains custom code which must be executed to
correctly load the dataset. You can inspect the repository content at
https://hf.co/datasets/conll2003.
You can avoid this prompt in future by passing the argument
`trust_remote_code=True`.

Do you wish to run the custom code? [y/N] y

Downloading data:   0%|          | 0.00/983k [00:00<?, ?B/s]

Generating train split:   0%|          | 0/14041 [00:00<?, ? examples/s]

Generating validation split:   0%|          | 0/3250 [00:00<?, ? examples/s]

Generating test split:   0%|          | 0/3453 [00:00<?, ? examples/s]
```

[5]:

```
from transformers import AutoTokenizer
from datasets import load_dataset

tok = AutoTokenizer.from_pretrained("bert-base-cased")

def tokenize_and_align_labels(examples):
    #
    tokenized = tok(
        examples["tokens"],
        truncation=True,
        is_split_into_words=True
    )

    aligned_labels = []
    for idx in range(len(examples["tokens"])):
        word_ids    = tokenized.word_ids(batch_index=idx)
        label_ids   = []
        previous_id = None

        for word_id in word_ids:
            if word_id is None:
                label_ids.append(-100)
            elif word_id != previous_id:
                label_ids.append(examples["ner_tags"][idx][word_id])
                previous_id = word_id
            else:
                label_ids.append(-100)

        aligned_labels.append(label_ids)

    tokenized["labels"] = aligned_labels
    return tokenized


conll = load_dataset("conll2003")
conll_enc = conll.map(tokenize_and_align_labels,
                      batched = True,
                      remove_columns = conll["train"].column_names)
conll_enc.set_format("torch")
```

```
tokenizer_config.json:   0%|          | 0.00/49.0 [00:00<?, ?B/s]

config.json:   0%|          | 0.00/570 [00:00<?, ?B/s]

vocab.txt:   0%|          | 0.00/213k [00:00<?, ?B/s]

tokenizer.json:   0%|          | 0.00/436k [00:00<?, ?B/s]

Map:   0%|          | 0/14041 [00:00<?, ? examples/s]
```

```
Map:   0%|          | 0/3250 [00:00<?, ? examples/s]

Map:   0%|          | 0/3453 [00:00<?, ? examples/s]
```

[6]:

```
!pip install seqeval
from transformers import DataCollatorForTokenClassification

data_collator = DataCollatorForTokenClassification(tokenizer=tok)  #       ␣
 ↪pad-
```

```
Collecting seqeval
  Downloading seqeval-1.2.2.tar.gz (43 kB)
                           43.6/43.6 kB
3.9 MB/s eta 0:00:00
  Installing build dependencies … done
  Getting requirements to build wheel … done
  Installing backend dependencies … done
  Preparing metadata (pyproject.toml) … done
Requirement already satisfied: numpy>=1.14.0 in /usr/local/lib/python3.11/dist-
packages (from seqeval) (2.0.2)
Requirement already satisfied: scikit-learn>=0.21.3 in
/usr/local/lib/python3.11/dist-packages (from seqeval) (1.6.1)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn>=0.21.3->seqeval) (1.15.2)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-
packages (from scikit-learn>=0.21.3->seqeval) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
/usr/local/lib/python3.11/dist-packages (from scikit-learn>=0.21.3->seqeval)
(3.6.0)
Building wheels for collected packages: seqeval
  Building wheel for seqeval (pyproject.toml) … done
  Created wheel for seqeval: filename=seqeval-1.2.2-py3-none-any.whl size=16249
sha256=f1eeaec8a5cbd2a0d6069f41ed479a5393bf98c7225b69774c87cc51966f79f0
  Stored in directory: /root/.cache/pip/wheels/bc/92/f0/243288f899c2eacdfa8c5f9a
ede4c71a9bad0ee26a01dc5ead
Successfully built seqeval
Installing collected packages: seqeval
Successfully installed seqeval-1.2.2
```

[7]:

```
from transformers import AutoModelForTokenClassification, TrainingArguments,␣
 ↪Trainer
import evaluate, torch
model = AutoModelForTokenClassification.from_pretrained(
    "bert-base-cased",
    num_labels=num_labels
)

args = TrainingArguments(
```

```python
    output_dir="ner_fast",
    per_device_train_batch_size=128,
    per_device_eval_batch_size=128,
    learning_rate=5e-5,
    logging_steps=50,
    num_train_epochs=1,
    weight_decay=0.01,
    fp16=torch.cuda.is_available()
)

metric = evaluate.load("seqeval")
def compute_metrics(pred):
    logits, labels = pred
    predictions = logits.argmax(-1)
    true, pred_tags = [], []
    for l, p in zip(labels, predictions):
        true.append([label_list[i] for i in l[l!=-100]])
        pred_tags.append([label_list[i] for i in p[l!=-100]])
    return metric.compute(predictions=pred_tags, references=true)

trainer = Trainer(
    model=model,
    args=args,
    train_dataset=conll_enc["train"],
    eval_dataset=conll_enc["validation"],
    compute_metrics=compute_metrics,
    data_collator=data_collator
)

trainer.train()
trainer.evaluate()
```

```
Xet Storage is enabled for this repo, but the 'hf_xet' package is not installed.
Falling back to regular HTTP download. For better performance, install the
package with: `pip install huggingface_hub[hf_xet]` or `pip install hf_xet`
WARNING:huggingface_hub.file_download:Xet Storage is enabled for this repo, but
the 'hf_xet' package is not installed. Falling back to regular HTTP download.
For better performance, install the package with: `pip install
huggingface_hub[hf_xet]` or `pip install hf_xet`
```

```
model.safetensors:   0%|          | 0.00/436M [00:00<?, ?B/s]
```

```
Some weights of BertForTokenClassification were not initialized from the model
checkpoint at bert-base-cased and are newly initialized: ['classifier.bias',
'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
WARNING:root:torch_xla.core.xla_model.xrt_world_size() will be removed in
```

```
release 2.7. is deprecated. Use torch_xla.runtime.world_size instead.
WARNING:root:torch_xla.core.xla_model.xla_model.get_ordinal() will be removed in
release 2.7. is deprecated. Use torch_xla.runtime.global_ordinal instead.
```

```
Downloading builder script:   0%|          | 0.00/6.34k [00:00<?, ?B/s]
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

[7]:

```
{'eval_loss': 0.05733289197087288,
 'eval_LOC': {'precision': 0.9189044038668098,
  'recall': 0.9314099074578116,
  'f1': 0.9251148959178156,
  'number': 1837},
 'eval_MISC': {'precision': 0.7603036876355749,
  'recall': 0.7603036876355749,
  'f1': 0.7603036876355749,
  'number': 922},
 'eval_ORG': {'precision': 0.8456090651558074,
  'recall': 0.8903803131991052,
  'f1': 0.8674173628768616,
  'number': 1341},
 'eval_PER': {'precision': 0.9608369098712446,
  'recall': 0.9723127035830619,
  'f1': 0.9665407447382622,
  'number': 1842},
 'eval_overall_precision': 0.8905940594059406,
 'eval_overall_recall': 0.908280040390441,
 'eval_overall_f1': 0.8993501083152808,
 'eval_overall_accuracy': 0.983762314551614,
 'eval_runtime': 186.5394,
 'eval_samples_per_second': 17.841,
 'eval_steps_per_second': 0.139,
 'epoch': 1.0}
```

[14]:

```python
from datasets import load_dataset, DatasetDict

raw = load_dataset("conll2003")         #         hub
#     ,          2000       ,
raw["train"] = raw["train"].shuffle(seed=228).select(range(2000))
ds = DatasetDict(train=raw["train"], validation=raw["validation"])
```

[15]:

```python

```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-15-b9462fbdb5c9> in <cell line: 0>()
     23
     24 # 2.
---> 25 ds_enc = ds.map(tokenize_and_align, batched=True,

     26                    remove_columns=ds["train"].column_names)
     27 ds_enc.set_format("torch")

NameError: name 'tokenize_and_align' is not defined
```

[17]:

```python
def read_conll(path):
    tokens, tags = [], []
    with open(path, encoding="utf8") as f:
        tok_buf, tag_buf = [], []
        for line in f:
            line = line.strip()
            if not line:
                if tok_buf:
                    tokens.append(tok_buf)
                    tags.append(tag_buf)
                    tok_buf, tag_buf = [], []
                continue
            token, tag = line.split()
            tok_buf.append(token)
            tag_buf.append(tag)
    if tok_buf:
        tokens.append(tok_buf)
        tags.append(tag_buf)
    return tokens, tags
from transformers import AutoTokenizer
tok = AutoTokenizer.from_pretrained("bert-base-cased")

def tokenize_and_align(batch):
    tok_out = tok(batch["tokens"],
                  is_split_into_words=True,
                  truncation=True)

    aligned_labels = []
    for i in range(len(batch["tokens"])):
        word_ids = tok_out.word_ids(batch_index=i)
        label_ids, prev = [], None
        for wid in word_ids:
            if wid is None:
                label_ids.append(-100)
            elif wid != prev:
                label_ids.append(batch["ner_tags"][i][wid])
                prev = wid
            else:
                label_ids.append(-100)
        aligned_labels.append(label_ids)

    tok_out["labels"] = aligned_labels
    return tok_out


ds_enc = ds.map(tokenize_and_align, batched=True,
                remove_columns=ds["train"].column_names)
ds_enc.set_format("torch")

trainer = Trainer(
    model=model,
    args=args,
    train_dataset=ds_enc["train"],
    eval_dataset=ds_enc["validation"],
    data_collator=data_collator,
    compute_metrics=compute_metrics,
)
trainer.train()
```

```
Map:   0%|          | 0/2000 [00:00<?, ? examples/s]

Map:   0%|          | 0/3250 [00:00<?, ? examples/s]

<IPython.core.display.HTML object>
```

[17]:

```
TrainOutput(global_step=16, training_loss=0.058396149426698685,
metrics={'train_runtime': 34.0088, 'train_samples_per_second': 60.22,
'train_steps_per_second': 0.47, 'total_flos': 71020053497952.0, 'train_loss':
0.058396149426698685, 'epoch': 1.0})
```

[19]:

```
label_feature = ds["train"].features["ner_tags"].feature
label_list    = label_feature.names
num_labels    = len(label_list)
id2tag        = {i: t for i, t in enumerate(label_list)}
```

[20]:

```
from transformers import AutoModelForTokenClassification

MODEL_NAME = "bert-base-cased"
model = AutoModelForTokenClassification.from_pretrained(
    MODEL_NAME,
    num_labels=num_labels
)
```

```
Some weights of BertForTokenClassification were not initialized from the model
checkpoint at bert-base-cased and are newly initialized: ['classifier.bias',
'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
```

[22]:

```python
tok = AutoTokenizer.from_pretrained(MODEL_NAME)

def tokenize_and_align(batch):
    tok_out = tok(batch["tokens"], is_split_into_words=True, truncation=True)
    new_labels = []
    for i in range(len(batch["tokens"])):
        word_ids = tok_out.word_ids(batch_index=i)
        label_ids, prev = [], None
        for wid in word_ids:
            if wid is None:
                label_ids.append(-100)
            elif wid != prev:
                label_ids.append(batch["ner_tags"][i][wid])
                prev = wid
            else:
                label_ids.append(-100)
        new_labels.append(label_ids)
    tok_out["labels"] = new_labels
    return tok_out

ds_enc = ds.map(tokenize_and_align, batched=True,
                remove_columns=ds["train"].column_names)
ds_enc.set_format("torch")
```

```
Map:   0%|          | 0/2000 [00:00<?, ? examples/s]

Map:   0%|          | 0/3250 [00:00<?, ? examples/s]
```

[23]:

```python
from transformers import DataCollatorForTokenClassification
collator = DataCollatorForTokenClassification(tok)
```

[31]:

```python
from transformers import TrainingArguments

args = TrainingArguments(
    output_dir     = "ner_custom_bert",
    num_train_epochs = 3,                # 1-3
    per_device_train_batch_size = 128,
    per_device_eval_batch_size  = 128,
    learning_rate = 5e-5,
    fp16          = torch.cuda.is_available(),  #         GPU
    logging_steps = 20,
    save_strategy = "epoch",             #
)
```

[32]:

```python
import evaluate, torch
seq_eval = evaluate.load("seqeval")

def compute_metrics(eval_pred):
    logits, labels = eval_pred
    preds = logits.argmax(-1)
    true, pred = [], []
    for y, p in zip(labels, preds):
        mask       = y != -100
        true.append([id2tag[i] for i in y[mask]])
        pred.append([id2tag[i] for i in p[mask]])
    return seq_eval.compute(predictions=pred, references=true)
```

[33]:

```
from transformers import Trainer

trainer = Trainer(
    model           = model,
    args            = args,
    train_dataset   = ds_enc["train"],
    eval_dataset    = ds_enc["validation"],
    data_collator   = collator,
    compute_metrics = compute_metrics,
)

trainer.train()
print(trainer.evaluate())
```

```
<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

{'eval_loss': 0.0883931815624237, 'eval_LOC': {'precision': 0.8247841543930929,
'recall': 0.8840500816548721, 'f1': 0.8533893851812926, 'number': 1837},
'eval_MISC': {'precision': 0.7383647798742138, 'recall': 0.6366594360086768,
'f1': 0.6837507280139778, 'number': 922}, 'eval_ORG': {'precision':
0.7270967741935483, 'recall': 0.8404175988068605, 'f1': 0.7796610169491525,
'number': 1341}, 'eval_PER': {'precision': 0.9539794260963725, 'recall':
0.9565689467969598, 'f1': 0.9552724315532666, 'number': 1842},
'eval_overall_precision': 0.8277876968024671, 'eval_overall_recall':
0.858296869740828, 'eval_overall_f1': 0.8427662563000908,
'eval_overall_accuracy': 0.975000973482341, 'eval_runtime': 5.0186,
'eval_samples_per_second': 663.133, 'eval_steps_per_second': 5.181, 'epoch':
3.0}
```

[34]:

```
# 7 task
from datasets import load_dataset
imdb = load_dataset("imdb")
```

```
README.md:   0%|          | 0.00/7.81k [00:00<?, ?B/s]

train-00000-of-00001.parquet:   0%|          | 0.00/21.0M [00:00<?, ?B/s]

test-00000-of-00001.parquet:   0%|          | 0.00/20.5M [00:00<?, ?B/s]
```

```
unsupervised-00000-of-00001.parquet:   0%|          | 0.00/42.0M [00:00<?, ?B/s]

Generating train split:   0%|          | 0/25000 [00:00<?, ? examples/s]

Generating test split:   0%|          | 0/25000 [00:00<?, ? examples/s]

Generating unsupervised split:   0%|          | 0/50000 [00:00<?, ? examples/s]
```

[35]:

```python
imdb["train"] = imdb["train"].shuffle(seed=42).select(range(10000))
imdb["test"]  = imdb["test"].shuffle(seed=42).select(range(5000))
```

[42]:

```python
from transformers import (AutoTokenizer, AutoModelForSequenceClassification,
                          TrainingArguments, Trainer, DataCollatorWithPadding)
import evaluate, torch

MODEL = "bert-base-uncased"
tok   = AutoTokenizer.from_pretrained(MODEL)

def tokenize(batch):
    return tok(batch["text"], truncation=True)

enc = imdb.map(tokenize, batched=True)
enc = enc.remove_columns(["text"])
enc.set_format("torch")

model = AutoModelForSequenceClassification.from_pretrained(MODEL, num_labels=2)
collator  = DataCollatorWithPadding(tok)
accuracy  = evaluate.load("accuracy")
f1        = evaluate.load("f1")

def compute_metrics(p):
    logits, labels = p
    preds = logits.argmax(-1)
    return {"acc": accuracy.compute(predictions=preds,␣
 ↪references=labels)["accuracy"],
            "f1":  f1.compute(predictions=preds, references=labels,␣
 ↪average="macro")["f1"]}

args = TrainingArguments(
    "bert_sentiment",
    num_train_epochs=2,
    per_device_train_batch_size=32,
    per_device_eval_batch_size=32,
    learning_rate=2e-5,
    fp16=torch.cuda.is_available(),
    save_strategy="no",
    logging_steps=100,
)
```

```
trainer = Trainer(model, args,
                  train_dataset=enc["train"],
                  eval_dataset =enc["test"],
                  data_collator=collator,
                  compute_metrics=compute_metrics)
trainer.train()
bert_metrics = trainer.evaluate()
```

```
Some weights of BertForSequenceClassification were not initialized from the
model checkpoint at bert-base-uncased and are newly initialized:
['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

[46]:

```
!pip install --upgrade torch torchvision torchaudio
!pip install --upgrade torchtext
```

```
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages
(2.6.0+cpu)
Collecting torch
  Downloading torch-2.7.0-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (29 kB)
Requirement already satisfied: torchvision in /usr/local/lib/python3.11/dist-
packages (0.21.0+cpu)
Collecting torchvision
  Downloading torchvision-0.22.0-cp311-cp311-manylinux_2_28_x86_64.whl.metadata
(6.1 kB)
Requirement already satisfied: torchaudio in /usr/local/lib/python3.11/dist-
packages (2.6.0+cpu)
Collecting torchaudio
  Downloading torchaudio-2.7.0-cp311-cp311-manylinux_2_28_x86_64.whl.metadata
(6.6 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from torch) (3.18.0)
Requirement already satisfied: typing-extensions>=4.10.0 in
/usr/local/lib/python3.11/dist-packages (from torch) (4.13.2)
Collecting sympy>=1.13.3 (from torch)
  Downloading sympy-1.14.0-py3-none-any.whl.metadata (12 kB)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-
packages (from torch) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages
(from torch) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages
(from torch) (2025.3.0)
Collecting nvidia-cuda-nvrtc-cu12==12.6.77 (from torch)
  Downloading nvidia_cuda_nvrtc_cu12-12.6.77-py3-none-
```

```
manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.6.77 (from torch)
  Downloading nvidia_cuda_runtime_cu12-12.6.77-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.6.80 (from torch)
  Downloading nvidia_cuda_cupti_cu12-12.6.80-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.5.1.17 (from torch)
  Downloading nvidia_cudnn_cu12-9.5.1.17-py3-none-
manylinux_2_28_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.6.4.1 (from torch)
  Downloading nvidia_cublas_cu12-12.6.4.1-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.3.0.4 (from torch)
  Downloading nvidia_cufft_cu12-11.3.0.4-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.7.77 (from torch)
  Downloading nvidia_curand_cu12-10.3.7.77-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.7.1.2 (from torch)
  Downloading nvidia_cusolver_cu12-11.7.1.2-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.5.4.2 (from torch)
  Downloading nvidia_cusparse_cu12-12.5.4.2-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparselt-cu12==0.6.3 (from torch)
  Downloading nvidia_cusparselt_cu12-0.6.3-py3-none-
manylinux2014_x86_64.whl.metadata (6.8 kB)
Collecting nvidia-nccl-cu12==2.26.2 (from torch)
  Downloading nvidia_nccl_cu12-2.26.2-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (2.0 kB)
Collecting nvidia-nvtx-cu12==12.6.77 (from torch)
  Downloading nvidia_nvtx_cu12-12.6.77-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-nvjitlink-cu12==12.6.85 (from torch)
  Downloading nvidia_nvjitlink_cu12-12.6.85-py3-none-
manylinux2010_x86_64.manylinux_2_12_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufile-cu12==1.11.1.6 (from torch)
  Downloading nvidia_cufile_cu12-1.11.1.6-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
Collecting triton==3.3.0 (from torch)
  Downloading triton-3.3.0-cp311-cp311-manylinux_2_27_x86_64.manylinux_2_28_x86_
64.whl.metadata (1.5 kB)
Requirement already satisfied: setuptools>=40.8.0 in
/usr/local/lib/python3.11/dist-packages (from triton==3.3.0->torch) (75.2.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages
(from torchvision) (2.0.2)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in
```

```
/usr/local/lib/python3.11/dist-packages (from torchvision) (11.2.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from sympy>=1.13.3->torch) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch) (3.0.2)
Downloading torch-2.7.0-cp311-cp311-manylinux_2_28_x86_64.whl (865.2 MB)
                         865.2/865.2 MB
1.8 MB/s eta 0:00:00
Downloading nvidia_cublas_cu12-12.6.4.1-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (393.1 MB)
                         393.1/393.1 MB
9.3 MB/s eta 0:00:00
Downloading nvidia_cuda_cupti_cu12-12.6.80-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (8.9 MB)
                         8.9/8.9 MB
168.8 MB/s eta 0:00:00
Downloading nvidia_cuda_nvrtc_cu12-12.6.77-py3-none-
manylinux2014_x86_64.whl (23.7 MB)
                         23.7/23.7 MB
130.8 MB/s eta 0:00:00
Downloading nvidia_cuda_runtime_cu12-12.6.77-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (897 kB)
                         897.7/897.7 kB
71.9 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-9.5.1.17-py3-none-manylinux_2_28_x86_64.whl
(571.0 MB)
                         571.0/571.0 MB
2.7 MB/s eta 0:00:00
Downloading nvidia_cufft_cu12-11.3.0.4-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (200.2 MB)
                         200.2/200.2 MB
16.9 MB/s eta 0:00:00
Downloading nvidia_cufile_cu12-1.11.1.6-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (1.1 MB)
                         1.1/1.1 MB
71.6 MB/s eta 0:00:00
Downloading nvidia_curand_cu12-10.3.7.77-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (56.3 MB)
                         56.3/56.3 MB
62.7 MB/s eta 0:00:00
Downloading nvidia_cusolver_cu12-11.7.1.2-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (158.2 MB)
                         158.2/158.2 MB
23.4 MB/s eta 0:00:00
Downloading nvidia_cusparse_cu12-12.5.4.2-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (216.6 MB)
                         216.6/216.6 MB
16.7 MB/s eta 0:00:00
```

```
Downloading nvidia_cusparselt_cu12-0.6.3-py3-none-manylinux2014_x86_64.whl
(156.8 MB)
                         156.8/156.8 MB
23.0 MB/s eta 0:00:00
Downloading nvidia_nccl_cu12-2.26.2-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (201.3 MB)
                         201.3/201.3 MB
4.0 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.6.85-py3-none-
manylinux2010_x86_64.manylinux_2_12_x86_64.whl (19.7 MB)
                         19.7/19.7 MB
137.7 MB/s eta 0:00:00
Downloading nvidia_nvtx_cu12-12.6.77-py3-none-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (89 kB)
                         89.3/89.3 kB
9.5 MB/s eta 0:00:00
Downloading
triton-3.3.0-cp311-cp311-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl (156.5
MB)
                         156.5/156.5 MB
24.1 MB/s eta 0:00:00
Downloading torchvision-0.22.0-cp311-cp311-manylinux_2_28_x86_64.whl (7.4
MB)
                         7.4/7.4 MB
157.8 MB/s eta 0:00:00
Downloading torchaudio-2.7.0-cp311-cp311-manylinux_2_28_x86_64.whl (3.5
MB)
                         3.5/3.5 MB
133.8 MB/s eta 0:00:00
Downloading sympy-1.14.0-py3-none-any.whl (6.3 MB)
                         6.3/6.3 MB
127.5 MB/s eta 0:00:00
Installing collected packages: nvidia-cusparselt-cu12, triton, sympy,
nvidia-nvtx-cu12, nvidia-nvjitlink-cu12, nvidia-nccl-cu12, nvidia-curand-cu12,
nvidia-cufile-cu12, nvidia-cuda-runtime-cu12, nvidia-cuda-nvrtc-cu12, nvidia-
cuda-cupti-cu12, nvidia-cublas-cu12, nvidia-cusparse-cu12, nvidia-cufft-cu12,
nvidia-cudnn-cu12, nvidia-cusolver-cu12, torch, torchvision, torchaudio
  Attempting uninstall: sympy
    Found existing installation: sympy 1.13.1
    Uninstalling sympy-1.13.1:
      Successfully uninstalled sympy-1.13.1
  Attempting uninstall: torch
    Found existing installation: torch 2.6.0+cpu
    Uninstalling torch-2.6.0+cpu:
      Successfully uninstalled torch-2.6.0+cpu
  Attempting uninstall: torchvision
    Found existing installation: torchvision 0.21.0+cpu
    Uninstalling torchvision-0.21.0+cpu:
      Successfully uninstalled torchvision-0.21.0+cpu
  Attempting uninstall: torchaudio
    Found existing installation: torchaudio 2.6.0+cpu
    Uninstalling torchaudio-2.6.0+cpu:
      Successfully uninstalled torchaudio-2.6.0+cpu
ERROR: pip's dependency resolver does not currently take into account all

the packages that are installed. This behaviour is the source of the following

dependency conflicts.

fastai 2.7.19 requires torch<2.7,>=1.10, but you have torch 2.7.0 which is

incompatible.

Successfully installed nvidia-cublas-cu12-12.6.4.1 nvidia-cuda-cupti-
cu12-12.6.80 nvidia-cuda-nvrtc-cu12-12.6.77 nvidia-cuda-runtime-cu12-12.6.77
nvidia-cudnn-cu12-9.5.1.17 nvidia-cufft-cu12-11.3.0.4 nvidia-cufile-
cu12-1.11.1.6 nvidia-curand-cu12-10.3.7.77 nvidia-cusolver-cu12-11.7.1.2 nvidia-
cusparse-cu12-12.5.4.2 nvidia-cusparselt-cu12-0.6.3 nvidia-nccl-cu12-2.26.2
nvidia-nvjitlink-cu12-12.6.85 nvidia-nvtx-cu12-12.6.77 sympy-1.14.0 torch-2.7.0
torchaudio-2.7.0 torchvision-0.22.0 triton-3.3.0

Requirement already satisfied: torchtext in /usr/local/lib/python3.11/dist-
packages (0.18.0)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages
(from torchtext) (4.67.1)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-
packages (from torchtext) (2.32.3)
Requirement already satisfied: torch>=2.3.0 in /usr/local/lib/python3.11/dist-
packages (from torchtext) (2.7.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages
(from torchtext) (2.0.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from torch>=2.3.0->torchtext) (3.18.0)
Requirement already satisfied: typing-extensions>=4.10.0 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext) (4.13.2)
Requirement already satisfied: sympy>=1.13.3 in /usr/local/lib/python3.11/dist-
packages (from torch>=2.3.0->torchtext) (1.14.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-
packages (from torch>=2.3.0->torchtext) (3.4.2)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages
(from torch>=2.3.0->torchtext) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages
(from torch>=2.3.0->torchtext) (2025.3.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.6.77 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext) (12.6.77)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.6.77 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext) (12.6.77)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.6.80 in
```

```
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext) (12.6.80)
Requirement already satisfied: nvidia-cudnn-cu12==9.5.1.17 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext)
(9.5.1.17)
Requirement already satisfied: nvidia-cublas-cu12==12.6.4.1 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext)
(12.6.4.1)
Requirement already satisfied: nvidia-cufft-cu12==11.3.0.4 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext)
(11.3.0.4)
Requirement already satisfied: nvidia-curand-cu12==10.3.7.77 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext)
(10.3.7.77)
Requirement already satisfied: nvidia-cusolver-cu12==11.7.1.2 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext)
(11.7.1.2)
Requirement already satisfied: nvidia-cusparse-cu12==12.5.4.2 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext)
(12.5.4.2)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.3 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext) (0.6.3)
Requirement already satisfied: nvidia-nccl-cu12==2.26.2 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext) (2.26.2)
Requirement already satisfied: nvidia-nvtx-cu12==12.6.77 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext) (12.6.77)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.6.85 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext) (12.6.85)
Requirement already satisfied: nvidia-cufile-cu12==1.11.1.6 in
/usr/local/lib/python3.11/dist-packages (from torch>=2.3.0->torchtext)
(1.11.1.6)
Requirement already satisfied: triton==3.3.0 in /usr/local/lib/python3.11/dist-
packages (from torch>=2.3.0->torchtext) (3.3.0)
Requirement already satisfied: setuptools>=40.8.0 in
/usr/local/lib/python3.11/dist-packages (from
triton==3.3.0->torch>=2.3.0->torchtext) (75.2.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests->torchtext) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests->torchtext) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests->torchtext) (2.4.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests->torchtext) (2025.4.26)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in
/usr/local/lib/python3.11/dist-packages (from
sympy>=1.13.3->torch>=2.3.0->torchtext) (1.3.0)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2->torch>=2.3.0->torchtext)
```

(3.0.2)

[58]:

```python
import torch, torch.nn as nn
from collections import Counter, defaultdict

counter = Counter()
for example in imdb["train"]:
    counter.update(example["text"].lower().split())

PAD, UNK = "<pad>", "<unk>"
vocab = {PAD: 0, UNK: 1}
for token, _ in counter.most_common():
    vocab[token] = len(vocab)

def encode_ids(text, max_len=256):
    ids = [vocab.get(tok, vocab[UNK]) for tok in text.lower().split()]
    return ids[:max_len]

def encode_example(example):
    example["input_ids"] = encode_ids(example["text"])
    return example

imdb_lstm = imdb.map(encode_example)
PAD_ID = vocab[PAD]

from torch.utils.data import DataLoader
MAX_LEN    = 256
PAD_ID     = vocab["<pad>"]

def yield_tokens(data_iter):
    for item in data_iter:
        yield item["text"].lower().split()

UNK_ID = vocab["<unk>"]
def encode_lstm(example):
    ids = [vocab.get(tok, UNK_ID)
           for tok in example["text"].lower().split()]
    example["input_ids"] = ids[:MAX_LEN]
    return example

imdb_lstm = imdb.map(encode_lstm)


def collate_fn(batch):
    seqs   = [b["input_ids"][:MAX_LEN] for b in batch]
    lens   = torch.tensor([len(s) for s in seqs])
    padded = torch.nn.utils.rnn.pad_sequence([torch.tensor(s) for s in seqs],
                                             batch_first=True,␣
 ↪padding_value=PAD_ID)
    labels = torch.tensor([b["label"] for b in batch])
    return padded, lens, labels

train_dl = DataLoader(imdb_lstm["train"], batch_size=64, shuffle=True,␣
 ↪collate_fn=collate_fn)
test_dl  = DataLoader(imdb_lstm["test"],  batch_size=64, shuffle=False,␣
 ↪collate_fn=collate_fn)

class SentLSTM(nn.Module):
    def __init__(self, vocab_size, embed_dim=128, hidden=128):
        super().__init__()
        self.emb   = nn.Embedding(vocab_size, embed_dim, padding_idx=PAD_ID)
        self.lstm  = nn.LSTM(embed_dim, hidden, batch_first=True)
        self.fc    = nn.Linear(hidden, 2)

    def forward(self, x, lens):
        x = self.emb(x)
        packed = nn.utils.rnn.pack_padded_sequence(x, lens.cpu(),␣
 ↪batch_first=True, enforce_sorted=False)
        _, (h, _) = self.lstm(packed)
        return self.fc(h[-1])

device = "cuda" if torch.cuda.is_available() else "cpu"
model_lstm = SentLSTM(len(vocab)).to(device)
opt  = torch.optim.AdamW(model_lstm.parameters(), lr=1e-3)
loss = nn.CrossEntropyLoss()

for epoch in range(3):
    model_lstm.train()
    for X, lens, y in train_dl:
        X, lens, y = X.to(device), lens.to(device), y.to(device)
        opt.zero_grad()
        out = model_lstm(X, lens)
        l   = loss(out, y)
        l.backward()
        opt.step()

from sklearn.metrics import accuracy_score, f1_score, confusion_matrix

model_lstm.eval()
all_preds, all_labels = [], []
with torch.no_grad():
    for X, lens, y in test_dl:
        X, lens = X.to(device), lens.to(device)
        logits  = model_lstm(X, lens)
        preds   = logits.argmax(-1).cpu()
        all_preds.extend(preds)
        all_labels.extend(y)

lstm_acc = accuracy_score(all_labels, all_preds)
lstm_f1  = f1_score(all_labels, all_preds, average="macro")
lstm_cm  = confusion_matrix(all_labels, all_preds)
```

[59]:
```
lstm_cm
```

[59]:
```
array([[1882,  612],
       [ 741, 1765]])
```

[60]:
```
lstm_f1
```

[60]:
```
0.7292517490877304
```

[61]:
```
lstm_acc
```

[61]:
```
0.7294
```

[63]:
```
bert_metrics
```

[63]:
```
{'eval_loss': 0.2011713683605194,
 'eval_acc': 0.9288,
 'eval_f1': 0.9287826685502184,
 'eval_runtime': 48.7342,
 'eval_samples_per_second': 103.09,
 'eval_steps_per_second': 3.222,
 'epoch': 2.0}
```

[66]:
```
!pip install tensorboard
```

```
Collecting tensorboard
  Downloading tensorboard-2.19.0-py3-none-any.whl.metadata (1.8 kB)
Requirement already satisfied: absl-py>=0.4 in /usr/local/lib/python3.11/dist-
packages (from tensorboard) (1.4.0)
Requirement already satisfied: grpcio>=1.48.2 in /usr/local/lib/python3.11/dist-
packages (from tensorboard) (1.71.0)
Requirement already satisfied: markdown>=2.6.8 in /usr/lib/python3/dist-packages
(from tensorboard) (3.3.6)
Requirement already satisfied: numpy>=1.12.0 in /usr/local/lib/python3.11/dist-
packages (from tensorboard) (2.0.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from tensorboard) (25.0)
Requirement already satisfied: protobuf!=4.24.0,>=3.19.6 in
/usr/local/lib/python3.11/dist-packages (from tensorboard) (5.29.4)
```

```
Requirement already satisfied: setuptools>=41.0.0 in
/usr/local/lib/python3.11/dist-packages (from tensorboard) (75.2.0)
Requirement already satisfied: six>1.9 in /usr/local/lib/python3.11/dist-
packages (from tensorboard) (1.17.0)
Collecting tensorboard-data-server<0.8.0,>=0.7.0 (from tensorboard)
  Downloading tensorboard_data_server-0.7.2-py3-none-
manylinux_2_31_x86_64.whl.metadata (1.1 kB)
Collecting werkzeug>=1.0.1 (from tensorboard)
  Downloading werkzeug-3.1.3-py3-none-any.whl.metadata (3.7 kB)
Requirement already satisfied: MarkupSafe>=2.1.1 in
/usr/local/lib/python3.11/dist-packages (from werkzeug>=1.0.1->tensorboard)
(3.0.2)
Downloading tensorboard-2.19.0-py3-none-any.whl (5.5 MB)
                         5.5/5.5 MB
100.5 MB/s eta 0:00:00
Downloading tensorboard_data_server-0.7.2-py3-none-
manylinux_2_31_x86_64.whl (6.6 MB)
                         6.6/6.6 MB
150.3 MB/s eta 0:00:00
Downloading werkzeug-3.1.3-py3-none-any.whl (224 kB)
                         224.5/224.5 kB
23.1 MB/s eta 0:00:00
Installing collected packages: werkzeug, tensorboard-data-server,
tensorboard
Successfully installed tensorboard-2.19.0 tensorboard-data-server-0.7.2
werkzeug-3.1.3
```

[73]:

```python
from datasets import load_dataset
from transformers import (AutoTokenizer, AutoModelForSequenceClassification,
                          TrainingArguments, Trainer, DataCollatorWithPadding)
import evaluate, torch

MAX_SAMPLES = 10000
MAX_LEN     = 128

ds = load_dataset("imdb")
ds["train"] = ds["train"].shuffle(seed=228).select(range(MAX_SAMPLES))

tok = AutoTokenizer.from_pretrained("distilbert-base-uncased")

def tok_fn(batch):
    return tok(batch["text"], truncation=True, max_length=MAX_LEN)

ds_tok = ds.map(tok_fn, batched=True).remove_columns(["text"])
ds_tok.set_format("torch")

num_labels = ds_tok["train"].features["label"].num_classes
```

```python
model = AutoModelForSequenceClassification.from_pretrained(
    "roberta-base", num_labels=num_labels)

args = TrainingArguments(
    "sentiment_fast",
    num_train_epochs=3,
    per_device_train_batch_size=128,
    per_device_eval_batch_size=128,
    learning_rate=3e-5,
    gradient_accumulation_steps=1,
    fp16=torch.cuda.is_available(),
    lr_scheduler_type="linear",
    warmup_ratio=0.1,
    save_strategy="no",
    logging_steps=50,
)

from transformers import (Trainer, DataCollatorWithPadding)
import evaluate

collator  = DataCollatorWithPadding(tok)
accuracy  = evaluate.load("accuracy")

def metrics(p):
    preds = p.predictions.argmax(-1)
    return {"acc": accuracy.compute(predictions=preds, references=p.
 ↪label_ids)["accuracy"]}

trainer = Trainer(model, args,
                  train_dataset=ds_tok["train"],
                  eval_dataset =ds_tok["test"],
                  data_collator=collator,
                  compute_metrics=metrics)

trainer.train()
print(trainer.evaluate())
```

```
Map:   0%|          | 0/10000 [00:00<?, ? examples/s]
```

```
Some weights of RobertaForSequenceClassification were not initialized from the
model checkpoint at roberta-base and are newly initialized:
['classifier.dense.bias', 'classifier.dense.weight', 'classifier.out_proj.bias',
'classifier.out_proj.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

```
{'eval_loss': 0.5800451040267944, 'eval_acc': 0.70496, 'eval_runtime': 9.6211,
'eval_samples_per_second': 2607.601, 'eval_steps_per_second': 20.372, 'epoch':
3.0}
```

[75]:

```
TensorFlow installation not found - running with reduced feature set.

NOTE: Using experimental fast data loading logic. To disable, pass
    "--load_fast=false" and report issues on GitHub. More details:
    https://github.com/tensorflow/tensorboard/issues/4784

Serving TensorBoard on localhost; to expose to the network, use a proxy or pass
--bind_all
TensorBoard 2.19.0 at http://localhost:6006/ (Press CTRL+C to quit)
```

[80]:

```python
from datasets import load_dataset
from transformers import (AutoTokenizer, AutoModelForTokenClassification,
                          DataCollatorForTokenClassification)

conll = load_dataset("conll2003")
label_list = conll["train"].features["ner_tags"].feature.names
tok = AutoTokenizer.from_pretrained("bert-base-cased")

def align(batch):
    tok_out = tok(batch["tokens"], is_split_into_words=True, truncation=True)
    new_labels = []
    for i in range(len(batch["tokens"])):
        word_ids = tok_out.word_ids(batch_index=i)
        ids, prev = [], None
        for wid in word_ids:
            if wid is None:
                ids.append(-100)
            elif wid != prev:
                ids.append(batch["ner_tags"][i][wid])
                prev = wid
            else:
                ids.append(-100)
        new_labels.append(ids)
    tok_out["labels"] = new_labels
    return tok_out

ds_tok = conll.map(align, batched=True,
                   remove_columns=conll["train"].column_names)
ds_tok.set_format("torch")

model = AutoModelForTokenClassification.from_pretrained(
    "bert-base-cased", num_labels=len(label_list))
collator = DataCollatorForTokenClassification(tok)
args = TrainingArguments(
    "bert_ner",
    num_train_epochs=3,
    per_device_train_batch_size=128,
    per_device_eval_batch_size=128,
    learning_rate=3e-5,
    gradient_accumulation_steps=1,

    fp16=torch.cuda.is_available(),
    lr_scheduler_type="linear",
    warmup_ratio=0.1,
    save_strategy="no",
    logging_steps=50,
)

trainer = Trainer(model, args,
                  train_dataset=ds_tok["train"],
                  eval_dataset=ds_tok["validation"],
                  data_collator=collator)
trainer.train()
print(trainer.evaluate())
```

```
Some weights of BertForTokenClassification were not initialized from the model
checkpoint at bert-base-cased and are newly initialized: ['classifier.bias',
'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
```

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

```
{'eval_loss': 0.04077420383691788, 'eval_runtime': 172.3002,
'eval_samples_per_second': 19.315, 'eval_steps_per_second': 0.151, 'epoch': 3.0}
```

```
[83]: %load_ext tensorboard
```

```
[84]: %tensorboard --logdir sentiment_fast/runs --host 0.0.0.0
```

```
<IPython.core.display.Javascript object>
```

```
[ ]:
```