Custom_Named_Entity_Recognition

September 24, 2023

[6]:
```
import json

with open("./drive/MyDrive/Dataset/Corona_NER/Corona2.json") as f:
  data = json.load(f)

print(data['examples'][0])
```

```
{'id': '18c2f619-f102-452f-ab81-d26f7e283ffe', 'content': "While bismuth
compounds (Pepto-Bismol) decreased the number of bowel movements in those with
travelers' diarrhea, they do not decrease the length of illness.[91] Anti-
motility agents like loperamide are also effective at reducing the number of
stools but not the duration of disease.[8] These agents should be used only if
bloody diarrhea is not present.[92]\n\nDiosmectite, a natural aluminomagnesium
silicate clay, is effective in alleviating symptoms of acute diarrhea in
children,[93] and also has some effects in chronic functional diarrhea,
radiation-induced diarrhea, and chemotherapy-induced diarrhea.[45] Another
absorbent agent used for the treatment of mild diarrhea is
kaopectate.\n\nRacecadotril an antisecretory medication may be used to treat
diarrhea in children and adults.[86] It has better tolerability than loperamide,
as it causes less constipation and flatulence.[94]", 'metadata': {},
'annotations': [{'id': '0825a1bf-6a6e-4fa2-be77-8d104701eaed', 'tag_id':
'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 371, 'start': 360, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'Medicine', 'value':
'Diosmectite', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:24:32.098000Z', 'annotator_id': 1, 'tagged_token_id':
'0825a1bf-6a6e-4fa2-be77-8d104701eaed', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'145f62b1-bbf5-42f1-8ad5-9c7e08337bf0', 'tag_id':
'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 408, 'start': 383, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'Medicine', 'value':
'aluminomagnesium silicate', 'correct': None, 'human_annotations':
[{'timestamp': '2020-03-21T00:24:43.692000Z', 'annotator_id': 1,
'tagged_token_id': '145f62b1-bbf5-42f1-8ad5-9c7e08337bf0', 'name': 'Ashpat123',
'reason': 'exploration'}], 'model_annotations': []}, {'id':
'243efeb2-723f-4be4-933c-fbbf7e0f9903', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 112, 'start': 104, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
'diarrhea', 'correct': None, 'human_annotations': [{'timestamp':
```

'2020-03-21T00:24:09.423000Z', 'annotator_id': 1, 'tagged_token_id':
'243efeb2-723f-4be4-933c-fbbf7e0f9903', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'281f49d3-879a-4409-b09e-4cfae019af16', 'tag_id':
'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 689, 'start': 679, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'Medicine', 'value':
'kaopectate', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:25:38.366000Z', 'annotator_id': 1, 'tagged_token_id':
'281f49d3-879a-4409-b09e-4cfae019af16', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'32fdf9e7-63f7-442a-8e25-f08ea4ad94d5', 'tag_id':
'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 23, 'start': 6, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'Medicine', 'value':
'bismuth compounds', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:23:46.721000Z', 'annotator_id': 1, 'tagged_token_id':
'32fdf9e7-63f7-442a-8e25-f08ea4ad94d5', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'392094d2-febf-4074-a2ca-4c0082f4e5b8', 'tag_id':
'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 37, 'start': 25, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'Medicine', 'value': 'Pepto-
Bismol', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:23:58.861000Z', 'annotator_id': 1, 'tagged_token_id':
'392094d2-febf-4074-a2ca-4c0082f4e5b8', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'450b8c30-cf2e-4774-96d9-58b4160bea38', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 470, 'start': 461, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
'diarrhea ', 'correct': None, 'human_annotations': [], 'model_annotations': []},
{'id': '6b73f683-7130-4e16-bcc2-e3cc8cf89f4d', 'tag_id':
'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 589, 'start': 577, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'Medicine', 'value':
'chemotherapy', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:25:24.179000Z', 'annotator_id': 1, 'tagged_token_id':
'6b73f683-7130-4e16-bcc2-e3cc8cf89f4d', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'74574b7f-d535-48e1-8651-2708adcfe453', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 865, 'start': 853, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
'constipation', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:26:31.676000Z', 'annotator_id': 1, 'tagged_token_id':
'74574b7f-d535-48e1-8651-2708adcfe453', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id': '7572ab8e-ae99-400c-b4ab-
ed46fbc9f97e', 'tag_id': 'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 198,
'start': 188, 'example_id': '18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name':
'Medicine', 'value': 'loperamide', 'correct': None, 'human_annotations':
[{'timestamp': '2020-03-21T00:24:17.680000Z', 'annotator_id': 1,
'tagged_token_id': '7572ab8e-ae99-400c-b4ab-ed46fbc9f97e', 'name': 'Ashpat123',
'reason': 'exploration'}], 'model_annotations': []}, {'id':

'800e6c6c-0bfb-4819-a25a-34f759753457', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 762, 'start': 754, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
'diarrhea', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:26:07.368000Z', 'annotator_id': 1, 'tagged_token_id':
'800e6c6c-0bfb-4819-a25a-34f759753457', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'8214556a-7584-4d9b-86cd-5e09137ad904', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 880, 'start': 870, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
'flatulence', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:26:41.322000Z', 'annotator_id': 1, 'tagged_token_id':
'8214556a-7584-4d9b-86cd-5e09137ad904', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'98968e14-6756-4174-9d3d-7abd58b3aa34', 'tag_id':
'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 833, 'start': 823, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'Medicine', 'value':
'loperamide', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:26:12.759000Z', 'annotator_id': 1, 'tagged_token_id':
'98968e14-6756-4174-9d3d-7abd58b3aa34', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'a0a03c7b-cfad-41ee-9f5c-f8a802475994', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 853, 'start': 852, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
' ', 'correct': None, 'human_annotations': [], 'model_annotations': []}, {'id':
'bfbddfd4-38aa-43a7-9366-24c95829ac8c', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 469, 'start': 461, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
'diarrhea', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:25:06.921000Z', 'annotator_id': 1, 'tagged_token_id':
'bfbddfd4-38aa-43a7-9366-24c95829ac8c', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'd7d68c18-0d8e-4547-a2fa-81fdcaf3080e', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 543, 'start': 535, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
'diarrhea', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:25:12.203000Z', 'annotator_id': 1, 'tagged_token_id':
'd7d68c18-0d8e-4547-a2fa-81fdcaf3080e', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'ee956220-42a4-4a91-b9f0-75019c4f5dd9', 'tag_id':
'c06bd022-6ded-44a5-8d90-f17685bb85a1', 'end': 704, 'start': 692, 'example_id':
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'Medicine', 'value':
'Racecadotril', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:25:56.503000Z', 'annotator_id': 1, 'tagged_token_id':
'ee956220-42a4-4a91-b9f0-75019c4f5dd9', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}, {'id':
'f04a6b7e-8904-405c-9301-e4543238b7a5', 'tag_id':
'03eb3e50-d4d8-4261-a60b-fa5aee5deb4a', 'end': 571, 'start': 563, 'example_id':

```
'18c2f619-f102-452f-ab81-d26f7e283ffe', 'tag_name': 'MedicalCondition', 'value':
'diarrhea', 'correct': None, 'human_annotations': [{'timestamp':
'2020-03-21T00:25:18.043000Z', 'annotator_id': 1, 'tagged_token_id':
'f04a6b7e-8904-405c-9301-e4543238b7a5', 'name': 'Ashpat123', 'reason':
'exploration'}], 'model_annotations': []}], 'classifications': []}
```

[11]:
```
training_data = {'classes' : ['MEDICINE', "MEDICALCONDITION", "PATHOGEN"],␣
 ↪'annotations' : []}

for example in data["examples"]:
  temp_dict = {}
  temp_dict["entities"] = []
  temp_dict["text"] = example["content"]


  for annotation in example["annotations"]:
    start = annotation["start"]
    end = annotation["end"]
    label = annotation["tag_name"].upper()
    temp_dict["entities"].append((start,end,label))
  training_data["annotations"].append(temp_dict)

print(training_data["annotations"][0])
```

```
{'entities': [(360, 371, 'MEDICINE'), (383, 408, 'MEDICINE'), (104, 112,
'MEDICALCONDITION'), (679, 689, 'MEDICINE'), (6, 23, 'MEDICINE'), (25, 37,
'MEDICINE'), (461, 470, 'MEDICALCONDITION'), (577, 589, 'MEDICINE'), (853, 865,
'MEDICALCONDITION'), (188, 198, 'MEDICINE'), (754, 762, 'MEDICALCONDITION'),
(870, 880, 'MEDICALCONDITION'), (823, 833, 'MEDICINE'), (852, 853,
'MEDICALCONDITION'), (461, 469, 'MEDICALCONDITION'), (535, 543,
'MEDICALCONDITION'), (692, 704, 'MEDICINE'), (563, 571, 'MEDICALCONDITION')],
'text': "While bismuth compounds (Pepto-Bismol) decreased the number of bowel
movements in those with travelers' diarrhea, they do not decrease the length of
illness.[91] Anti-motility agents like loperamide are also effective at reducing
the number of stools but not the duration of disease.[8] These agents should be
used only if bloody diarrhea is not present.[92]\n\nDiosmectite, a natural
aluminomagnesium silicate clay, is effective in alleviating symptoms of acute
diarrhea in children,[93] and also has some effects in chronic functional
diarrhea, radiation-induced diarrhea, and chemotherapy-induced diarrhea.[45]
Another absorbent agent used for the treatment of mild diarrhea is
kaopectate.\n\nRacecadotril an antisecretory medication may be used to treat
diarrhea in children and adults.[86] It has better tolerability than loperamide,
as it causes less constipation and flatulence.[94]"}
```

[12]:
```
import spacy
from spacy.tokens import DocBin
from tqdm import tqdm
```

```
nlp = spacy.blank("en") # load a new spacy model
doc_bin = DocBin() # create a DocBin object to store docs
```

[13]:
```
from spacy.util import filter_spans

for training_example in tqdm(training_data["annotations"]):
  text = training_example["text"]
  labels = training_example["entities"]
  doc = nlp.make_doc(text)
  ents = []

  for start, end, label in labels:
    span = doc.char_span(start, end, label=label, alignment_mode= "contract")

    if span is None:
      print("Skipping entity ")
    else:
      ents.append(span)
  filtered_ents = filter_spans(ents) # There might be a chance of indices of␣
 ↪some entities to overlap, hence we use filter_spans to filter the entities
  doc.ents = filtered_ents
  doc_bin.add(doc)

doc_bin.to_disk("training_data.spacy") # save the training data
```

```
100%|      | 31/31 [00:00<00:00, 318.54it/s]
```

```
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
Skipping entity
```

[21]:
```
!python -m spacy init fill-config base_config.cfg config.cfg
```

```
2023-09-24 03:39:39.543910: W
```

```
tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not
find TensorRT
  Auto-filled config with all values
  Saved config
config.cfg
You can now add your data and train your pipeline:
python -m spacy train config.cfg --paths.train ./train.spacy --paths.dev
./dev.spacy
```

[22]:
```
!python -m spacy train config.cfg --output ./ --paths.train ./training_data.
↪spacy --paths.dev ./training_data.spacy
```

```
2023-09-24 03:39:56.405801: W
tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not
find TensorRT
  Saving to output directory: .
  Using CPU


=========================== Initializing pipeline

===========================
  Initialized pipeline


============================== Training pipeline

==============================
  Pipeline: ['tok2vec', 'ner']
  Initial learn rate: 0.001
```

E	#	LOSS TOK2VEC	LOSS NER	ENTS_F	ENTS_P	ENTS_R	SCORE
0	0	0.00	153.29	0.00	0.00	0.00	0.00
7	200	734.85	3155.35	76.73	79.66	74.02	0.77
14	400	669.55	734.48	97.24	97.24	97.24	0.97
22	600	945.17	272.49	98.03	98.03	98.03	0.98
30	800	188.78	219.05	98.23	98.04	98.43	0.98
40	1000	224.41	192.51	98.41	99.20	97.64	0.98
51	1200	206.75	185.68	98.24	97.67	98.82	0.98
65	1400	142.57	143.77	98.43	98.43	98.43	0.98
81	1600	132.46	167.75	98.23	98.04	98.43	0.98
101	1800	3742.49	248.50	98.43	98.05	98.82	0.98
127	2000	121.44	217.56	98.43	98.43	98.43	0.98
159	2200	177.33	245.35	98.62	98.43	98.82	0.99
198	2400	421.83	260.98	98.82	98.44	99.21	0.99
246	2600	97.26	269.22	98.81	99.21	98.43	0.99
295	2800	140.40	277.83	98.82	98.44	99.21	0.99
345	3000	120.68	279.61	98.82	98.44	99.21	0.99
393	3200	174.15	268.27	98.82	98.82	98.82	0.99
443	3400	138.91	268.67	98.81	99.21	98.43	0.99

```
492     3600         533.68    259.68    98.81    99.21    98.43    0.99
541     3800         110.02    253.92    98.81    99.21    98.43    0.99
590     4000          55.16    243.94    98.81    99.21    98.43    0.99
 Saved pipeline to output directory
model-last
```

[23]:
```
nlp_ner = spacy.load("model-best")

doc = nlp_ner("Antiretroviral therapy (ART) is recommended for all HIV-infected\
individuals to reduce the risk of disease progression.\nART also is recommended␣
 ↪\
for HIV-infected individuals for the prevention of transmission of HIV.
 ↪\nPatients \
starting ART should be willing and able to commit to treatment and understand␣
 ↪the\
benefits and risks of therapy and the importance of adherence. Patients may␣
 ↪choose\
to postpone therapy, and providers, on a case-by-case basis, may elect to defer\
therapy on the basis of clinical and/or psychosocial factors.")

colors = {"PATHOGEN": "#F67DE3", "MEDICINE": "#7DF6D9", "MEDICALCONDITION":
 ↪"#FFFFFF"}
options = {"colors": colors}

spacy.displacy.render(doc, style="ent", options= options, jupyter=True)
```

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

[ ]: