Search models, datasets, users...

Datasets: eriktks/conll2003

♡ like 123

Tasks: Token Classification Sub-tasks: named-entity-recognition part-of-speech

Languages: English Size: 10K<n<100K License: other

Dataset card Files Community 12

Downloads last month 22,155

Edit dataset card

Papers with Code

Homepage:
aclweb.org

Size of downloaded dataset files:
4.85 MB

Models trained or fine-tuned on `eriktks/conll2003`

`dslim/bert-base-NER`
Token Classification • Updated 24 days ago • 2.38M • ♡ 513

`flair/ner-english-fast`
Token Classification • Updated Jul 21 • 1.21M • ♡ 20

`flair/ner-english-large`
Token Classification • Updated May 8, 2021 • 579k • ♡ 43

`flair/ner-english`
Token Classification • Updated Jul 21 • 415k • ♡ 30

Browse 995 models trained on this dataset

Spaces using eriktks/conll2003 8

KarishmaShirsath/PIIMasking   bhavanishankarpullela/CoSTA   sonic314/choose-your-transformer   XS5217/text-classification   187Matt/RainbowSpace   + 3 Spaces

Dataset Viewer   Full Screen Viewer

The viewer is disabled because this dataset repo requires arbitrary Python code execution. Please consider removing the loading script and relying on automated data support (you can use convert_to_parquet from the datasets library). If this is not possible, please open a discussion for direct help.

Dataset Card for "conll2003"

Dataset Summary

The shared task of CoNLL-2003 concerns language-independent named entity recognition. We will concentrate on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups.

The CoNLL-2003 shared task data files contain four columns separated by a single space. Each word has been put on a separate line and there is an empty line after each sentence. The first item on each line is a word, the second a part-of-speech (POS) tag, the third a syntactic chunk tag and the fourth the named entity tag. The chunk tags and the named entity tags have the format I-TYPE which means that the word is inside a phrase of type TYPE. Only if two phrases of the same type immediately follow each other, the first word of the second phrase will have tag B-TYPE to show that it starts a new phrase. A word with tag O is not part of a phrase. Note the dataset uses IOB2 tagging scheme, whereas the original dataset uses IOB1.

For more details see https://www.clips.uantwerpen.be/conll2003/ner/ and https://www.aclweb.org/anthology/W03-0419

Supported Tasks and Leaderboards

More Information Needed

Languages

More Information Needed

Dataset Structure

Data Instances

conll2003

- **Size of downloaded dataset files:** 4.85 MB
- **Size of the generated dataset:** 10.26 MB
- **Total amount of disk used:** 15.11 MB

An example of 'train' looks as follows.

```
{
    "chunk_tags": [11, 12, 12, 21, 13, 11, 11, 21, 13, 11, 12, 13, 1
    "id": "0",
    "ner_tags": [0, 3, 4, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 7, 0,
    "pos_tags": [12, 22, 22, 38, 15, 22, 28, 38, 15, 16, 21, 35, 24,
    "tokens": ["The", "European", "Commission", "said", "on", "Thurs
}
```

The original data files have `-DOCSTART-` lines used to separate documents, but these lines are removed here. Indeed `-DOCSTART-` is a special line that acts as a boundary between two different documents, and it is filtered out in this implementation.

Data Fields

The data fields are the same among all splits.

conll2003

- `id`: a `string` feature.
- `tokens`: a `list` of `string` features.
- `pos_tags`: a `list` of classification labels (`int`). Full tagset with indices:

```
{'"': 0, "''": 1, '#': 2, '$': 3, '(': 4, ')': 5, ',': 6, '.': 7, ':
 'EX': 13, 'FW': 14, 'IN': 15, 'JJ': 16, 'JJR': 17, 'JJS': 18, 'LS':
 'NNS': 24, 'NN|SYM': 25, 'PDT': 26, 'POS': 27, 'PRP': 28, 'PRP$': 2
 'SYM': 34, 'TO': 35, 'UH': 36, 'VB': 37, 'VBD': 38, 'VBG': 39, 'VBN
 'WP': 44, 'WP$': 45, 'WRB': 46}
```

- `chunk_tags`: a `list` of classification labels (`int`). Full tagset with indices:

```
{'O': 0, 'B-ADJP': 1, 'I-ADJP': 2, 'B-ADVP': 3, 'I-ADVP': 4, 'B-CONJ
 'B-LST': 9, 'I-LST': 10, 'B-NP': 11, 'I-NP': 12, 'B-PP': 13, 'I-PP'
 'I-SBAR': 18, 'B-UCP': 19, 'I-UCP': 20, 'B-VP': 21, 'I-VP': 22}
```

- `ner_tags`: a `list` of classification labels (`int`). Full tagset with indices:

```
{'O': 0, 'B-PER': 1, 'I-PER': 2, 'B-ORG': 3, 'I-ORG': 4, 'B-LOC': 5,
```

Data Splits

name	train	validation	test

name	train	validation	test
conll2003	14041	3250	3453

Dataset Creation

Curation Rationale

More Information Needed

Source Data

Initial Data Collection and Normalization

More Information Needed

Who are the source language producers?

More Information Needed

Annotations

Annotation process

More Information Needed

Who are the annotators?

More Information Needed

Personal and Sensitive Information

More Information Needed

Considerations for Using the Data

Social Impact of Dataset

More Information Needed

Discussion of Biases

More Information Needed

Other Known Limitations

More Information Needed

Additional Information

Dataset Curators

More Information Needed

Licensing Information

From the CoNLL2003 shared task page:

> *“The English data is a collection of news wire articles from the Reuters Corpus. The annotation has been done by people of the University of Antwerp. Because of copyright reasons we only make available the annotations. In order to build the complete data sets you will need access to the Reuters Corpus. It can be obtained for research purposes without any charge from NIST.”*

The copyrights are defined below, from the Reuters Corpus page:

> *“The stories in the Reuters Corpus are under the copyright of Reuters Ltd and/or Thomson Reuters, and their use is governed by the following agreements:*
>
> *Organizational agreement*

> *This agreement must be signed by the person responsible for the data at your organization, and sent to NIST.*
>
> *Individual agreement*
>
> *This agreement must be signed by all researchers using the Reuters Corpus at your organization, and kept on file at your organization.”*

Citation Information

```
@inproceedings{tjong-kim-sang-de-meulder-2003-introduction,
    title = "Introduction to the {C}o{NLL}-2003 Shared Task: Languag
    author = "Tjong Kim Sang, Erik F.  and
      De Meulder, Fien",
    booktitle = "Proceedings of the Seventh Conference on Natural La
```

```
    year = "2003",
    url = "https://www.aclweb.org/anthology/W03-0419",
    pages = "142--147",
}
```

Contributions

Thanks to @jplu, @vblagoje, @lhoestq for adding this dataset.