NER SpaCy German

Description

Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Most research on NER systems has been structured as taking an unannotated block of text, such as this one:

Jim bought 300 shares of Acme Corp. in 2006.

And producing an annotated block of text that highlights the names of entities:

$[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$ in $[\text{2006}]_{\text{Time}}$.

In this example, a person name consisting of one token, a two-token company name and a temporal expression have been detected and classified.

In case of German SpaCy NER for the example input sentence:

```
GER: Es    schneit in New York.
EN:  It is snowing in New York.
```

the output is:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE chunkList SYSTEM "ccl.dtd">
<chunkList>
 <chunk>
  <sentence>
   <tok>
    <orth>Es</orth>
    <lex disamb="1"><base>es</base><ctag>PRON</ctag></lex>
    <ann chan="LOC">0</ann>
   </tok>
   <tok>
    <orth>schneit</orth>
    <lex disamb="1"><base>Schneit</base><ctag>ADJ</ctag></lex>
    <ann chan="LOC">0</ann>
   </tok>
   <tok>
    <orth>in</orth>
    <lex disamb="1"><base>in</base><ctag>ADP</ctag></lex>
    <ann chan="LOC">0</ann>
   </tok>
   <tok>
```

```
    <orth>New</orth>
    <lex disamb="1"><base>New</base><ctag>PROPN</ctag></lex>
    <ann chan="LOC">1</ann>
   </tok>
   <tok>
    <orth>York</orth>
    <lex disamb="1"><base>York</base><ctag>PROPN</ctag></lex>
    <ann chan="LOC">1</ann>
   </tok>
   <ns/>
   <tok>
    <orth>.</orth>
    <lex disamb="1"><base>.</base><ctag>PUNCT</ctag></lex>
    <ann chan="LOC">0</ann>
   </tok>
  </sentence>
 </chunk>
</chunkList>
```

The information about the recognised named entity is stored within `<ann></ann>` section:

```
   <tok>
    <orth>New</orth>
    <lex disamb="1"><base>New</base><ctag>PROPN</ctag></lex>
    <ann chan="LOC">1</ann>
   </tok>
   <tok>
    <orth>York</orth>
    <lex disamb="1"><base>York</base><ctag>PROPN</ctag></lex>
    <ann chan="LOC">1</ann>
   </tok>
```

The format of annotation information at the level of token is:
`<ann chan=”annotation_category”>annotation_number</ann>`
All tokens with the same `annotation_category` and `annotation_number` belong to the same annotation, in this case [New York]$_{LOC}$
LOC - non-GPE locations, mountain ranges, bodies of water. See the full list of categories in **Output** section.

Input

Plain text file (UTF-8) in German.

Output

File in CCL format. Categories of named entities stored as:

`<ann chan=“category_name”>annotation_number</ann>`

are described here. Categories:

- PERSON People, including fictional.

- NORP Nationalities or religious or political groups.
- FAC Buildings, airports, highways, bridges, etc.
- ORG Companies, agencies, institutions, etc.
- GPE Countries, cities, states.
- LOC Non-GPE locations, mountain ranges, bodies of water.
- PRODUCT Objects, vehicles, foods, etc. (Not services.)
- EVENT Named hurricanes, battles, wars, sports events, etc.
- WORK_OF_ART Titles of books, songs, etc.
- LAW Named documents made into laws.
- LANGUAGE Any named language.
- DATE Absolute or relative dates or periods.
- TIME Times smaller than a day.
- PERCENT Percentage, including "%".
- MONEY Monetary values, including unit.
- QUANTITY Measurements, as of weight or distance.
- ORDINAL "first", "second", etc.
- CARDINAL Numerals that do not fall under another type.