

NER Liner2 Polish

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:

[Jim]_{Person} bought 300 shares of [Acme Corp.]_{Organization} in [2006]_{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 Polish Liner2 NER for the example input sentence:

PL: W Nowym Jorku pada śnieg.

EN: It is snowing in New York.

the output is:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE chunkList SYSTEM "ccl.dtd">
<chunkList>
  <chunk type="p" id="ch1">
    <sentence id="s1">
      <tok>
        <orth>W</orth>
        <lex disamb="1"><base>w</base><ctag>prep:acc:nwok</ctag></lex>
        <ann chan="nam_loc">0</ann>
      </tok>
      <tok>
        <orth>Nowym</orth>
        <lex disamb="1"><base>nowa</base><ctag>subst:pl:dat:f</ctag></lex>
        <ann chan="nam_loc">1</ann>
      </tok>
      <tok>
        <orth>Jorku</orth>
        <lex disamb="1"><base>Jork</base><ctag>subst:sg:gen:m3</ctag></lex>
        <ann chan="nam_loc">1</ann>
      </tok>
      <tok>
```

```

    <orth>pada</orth>
    <lex disamb="1"><base>padać</base><ctag>fin:sg:ter:imperf</ctag></lex>
    <ann chan="nam_loc">0</ann>
</tok>
<tok>
    <orth>śnieg</orth>
    <lex disamb="1"><base>śnieg</base><ctag>subst:sg:nom:m3</ctag></lex>
    <ann chan="nam_loc">0</ann>
</tok>
<ns/>
<tok>
    <orth>.</orth>
    <lex disamb="1"><base>.</base><ctag>interp</ctag></lex>
    <ann chan="nam_loc">0</ann>
</tok>
</sentence>
</chunk>
</chunkList>

```

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

```

<tok>
    <orth>Nowym</orth>
    <lex disamb="1"><base>nowa</base><ctag>subst:pl:dat:f</ctag></lex>
    <ann chan="nam_loc">1</ann>
</tok>
<tok>
    <orth>Jorku</orth>
    <lex disamb="1"><base>Jork</base><ctag>subst:sg:gen:m3</ctag></lex>
    <ann chan="nam_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 [Nowym Jorku]_{nam_loc}

`nam_loc` is **location** - names of geographical (e.g. mountains, rivers) and geopolitical entities (e.g., countries, cities). See the full list of categories in **Output** section.

Input

[Plain text file \(UTF-8\)](#) in Polish.

Output

File in [CCL](#) format. Categories of named entities stored as:

```
<ann chan="category_name">annotation_number</ann>
```

are described in [this](#) article. Categories:

- event, `nam_eve` – names of events organized by humans,
- facility – names of buildings and stationary constructions (e.g. monuments) developed by humans,
- living, `nam_liv` – people names,

- location, `nam_loc` – names of geographical (e.g, mountains, rivers) and geopolitical entities (e.g., countries, cities),
- organization, `nam_org` – names of organizations, institutions, organized groups of people,
- product, `nam_pro` – names of artifacts created or manufactured by humans (products of mass production, arts, books, newspapers, etc.),
- adjective, `nam_adj` – adjective forms of proper names,
- numerical, `nam_num` – numerical identifiers which indicate entities,
- other, `nam_oth` – other names which do not fit into previous categories.