Comparing Annotated Datasets for Named Entity Recognition in English Literature

Rositsa V. Ivanova\(^1\), Sabrina Kirrane\(^1\), and Marieke van Erp\(^2\)

\(^1\) Vienna University of Economics and Business, \(^2\) KNAW Humanities Cluster

Objective

Over the years, researchers and engineers have investigated various approaches that could potentially improve the recognition and linking of named entities. The growing interest in named entity recognition (NER) in various domains has led to the creation of different benchmark datasets. We take a closer look at existing annotated NER datasets in the domain of English literature and compare the performance of NER tools using such annotated datasets as a means to detect the differences between the datasets.

Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LitBank</th>
<th>Flair</th>
<th>OWTO</th>
<th>New CoNLL</th>
<th>New (Extended)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No.</td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Novels</td>
<td>100</td>
<td>48</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Annotators</td>
<td>95 by one annotator, 10 by two annotators</td>
<td>one</td>
<td>one</td>
<td>one</td>
<td>one</td>
</tr>
<tr>
<td>Inter-ann. F</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>Annot. layers</td>
<td>multiple</td>
<td>one</td>
<td>one</td>
<td>one</td>
<td>one</td>
</tr>
</tbody>
</table>

Table 1: Characteristics of datasets

Tools

- BookNLP (Bamman et al., 2014) - targeting the domain of English novels
- Flair (Akbik et al., 2019; Schweter and Akbik, 2020) - one of the best performing NER tools

Analysis

- The main difference between the performance comes down to honorifics.
- In some novels (e.g. Ulysses) personal pronouns make up to 10% of all tokens, leading to a drop in the recall whenever tools do not tag them.
- Inconsistencies in the gold standards (e.g. same entity tagged differently throughout the novel) lead to drops in the precision score of tools.
- Handling of common phrases (e.g. a boy) is done differently in the various gold standards, however neither of the tools tags them. This leads to further decrease in the recall values of the tools.

Discussion

Using existing gold standards:
- + datasets already exist
- + adaptation of existing annotations is possible
- - literary texts differ from other types of texts (e.g. news)
- - the annotation has been made following specific guidelines
- Dataset maintenance and using old datasets:
- + evaluation is comparable over the years and easy to execute
- + very limited view on the problem of NER
- - tools are adapted to perform better with the datasets

Evaluation Metrics

- + open evaluation over the years and easy to execute
- + only accepts full matches as correct
- - missing differentiation between “ambiguous” and “incorrect” tags
- - small differences in the definition of an entity as improved by the frequency of its occurrence

Annotation and Training Challenges

- work could be invested into recognizing shortcomings of existing datasets
- - insufficient annotation guidelines for inexperienced annotators
- - literary texts differ from other types of texts (e.g. news)

Selected References


Contact Information

- n.ivanova@wu.ac.at
- https://github.com/therosko/annotated_datasets_en_comparisson

Selected References


