Beyond Metadata: What Paper Authors Say About Corpora They Use

Nikolay Kolyada    Martin Potthast    Benno Stein

Bauhaus-Universität Weimar
Leipzig University
webis.de
Introduction

- How do we search for datasets?
  - Web-search
  - Data repositories: CKAN, Socrata, OpenDataSoft
  - DCAT, schema.org/Dataset
Introduction

- How do we search for datasets?
  - Web-search
  - Data repositories: CKAN, Socrata, OpenDataSoft
  - DCAT, schema.org/Dataset

- Google Dataset Search
  - Indexes the metadata contained in the web-pages
  - Provides search over the metadata
  - Improves search for the datasets in the long-tail
Introduction

- How do we search for datasets?
  - Web-search
  - Data repositories: CKAN, Socrata, OpenDataSoft
  - DCAT, schema.org/Dataset

- Google Dataset Search
  - Indexes the metadata contained in the web-pages
  - Provides search over the metadata
  - Improves search for the datasets in the long-tail"

- What is the problem?
  - The dataset search relies exclusively on metadata provided by publishers which quality varies significantly
  - The provided metadata says little about the underlying properties of a dataset
Introduction

- Dataset user experience is missing

![Customer Ratings](image)

- Tomatometer: 77%
- Audience Score: 66%

![Certification](image)

- Total Count: 57
- User Ratings: 124
Collecting Metadata for NLP datasets

- Parsing the catalogs
  - Pre-process the metadata, remove duplicates, resolve ambiguities
  - Normalize properties and format accordingly to schema.org/Dataset

<table>
<thead>
<tr>
<th>Resource</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>European Language Association (ELRA)</td>
<td>5,398</td>
</tr>
<tr>
<td>Linguistic Data Consortium (LDC)</td>
<td>950</td>
</tr>
<tr>
<td>Language Resources monitoring (LRE Map)</td>
<td>6,143</td>
</tr>
<tr>
<td>NLP Progress</td>
<td>90</td>
</tr>
<tr>
<td>Big Bad NLP Database</td>
<td>791</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td><strong>13,372</strong></td>
</tr>
</tbody>
</table>

- Datasets

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper or authors info</td>
<td>7,983</td>
</tr>
<tr>
<td>Unique dataset names</td>
<td>10,445</td>
</tr>
</tbody>
</table>
Dataset Mentions

- Mining from ACL Anthology

ACL Anthology articles → PDF to text → - pre-processing - reference, footnote resolution - coreference resolution → - dataset mentions extraction
Dataset Mentions

- Mining from ACL Anthology

- Extracted mentions

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Number of elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>separate mentions</td>
<td>466,567</td>
</tr>
<tr>
<td>publications with at least one mention</td>
<td>53,129</td>
</tr>
<tr>
<td>unique datasets mentioned</td>
<td>(22%) 2,986</td>
</tr>
<tr>
<td>coreference cases</td>
<td>93,176</td>
</tr>
</tbody>
</table>

ACL Anthology articles → PDF to text → - pre-processing, reference, footnote resolution, coreference resolution → - dataset mentions extraction
Dataset Mentions

- Taxonomy of Dataset Mentions

Description

“For instance, two words are said to be synonyms if they belong in the same synset in the WordNet.”

Reuse

“Second, we reuse the RCV1-V2, using a version that contained a selected 5,000 term vocabulary.”

Personal experience

“When only WordNet, not BabelNet, is used for identifying lexico-semantic relations, performance increases slightly, which we attribute to noise that comes with using BabelNet.”
Conclusion

- Corpus of NLP datasets
- Corpus of dataset mentions in ACL Anthology
- Taxonomy of dataset mentions
- Future steps
  - Evaluation of the mention extraction approaches.
  - Scaling up to different fields.

https://webis.de/data/webis-dataset-reviews-21.html