Beyond Metadata: What Paper Authors Say About Corpora They Use

Nikolay Kolyada Martin Potthast Benno Stein

Bauhaus-Universität Weimar

Leipzig University

webis.de

The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing

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

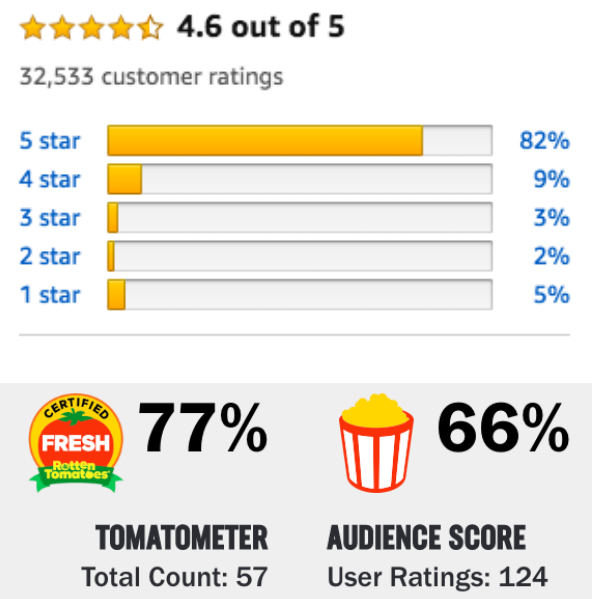
- How do we search for datasets?
  - Web-search
  - Data repositories: CKAN, Socrata, OpenDataSoft
  - DCAT, schema.org/Dataset

- 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



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

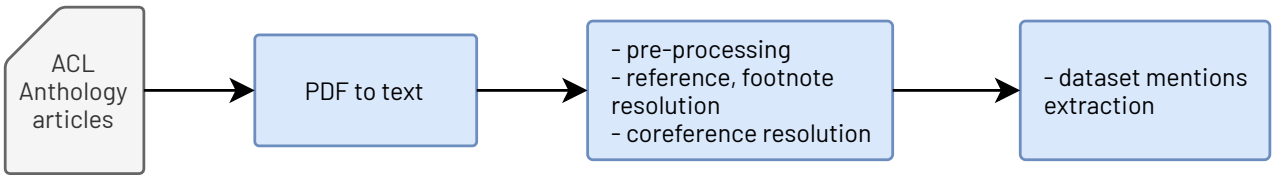
Resource	Datasets
European Language Association (ELRA)	5,398
Linguistic Data Consortium (LDC)	950
Language Resources monitoring (LRE Map)	6,143
NLP Progress	90
Big Bad NLP Database	791
Σ	13,372

- Datasets

Characteristic	Datasets
Paper or authors info	7,983
Unique dataset names	10,445

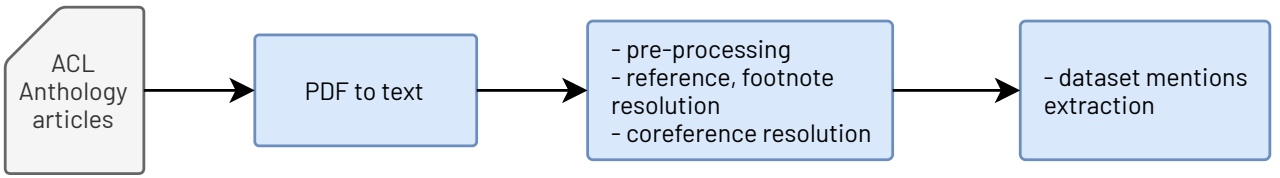
Dataset Mentions

- Mining from ACL Anthology



Dataset Mentions

- Mining from ACL Anthology

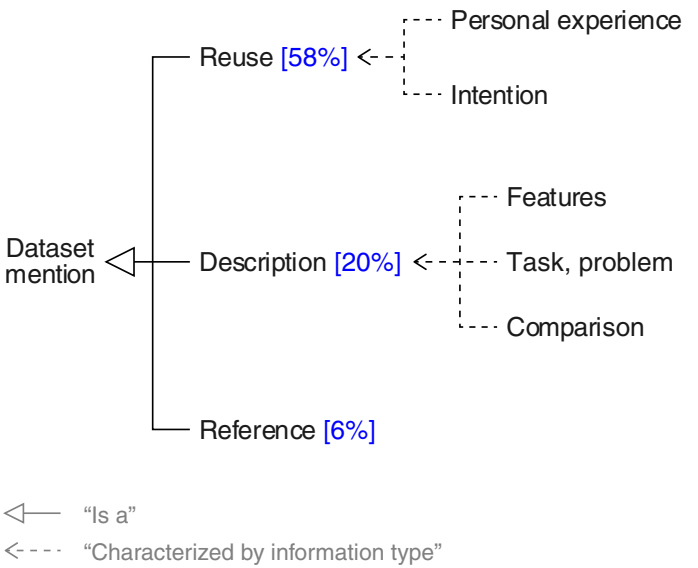


- Extracted mentions

Characteristic	Number of elements
separate mentions	466,567
publications with at least one mention	53,129
unique datasets mentioned	(22%) 2,986
coreference cases	93,176

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