Datasheet for dataset: From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains

I. Dataset description

This dataset is based on the 1641 depositions as annotated by [1] and the Women Writers Online corpus (WWO) by [2]. It contains texts of Early Modern English from various genres like law, novels, or poems from 1600 to around 1900. It is used for entity linking. We use the personography of WWO to create a knowledge base against which named entities in the WWO corpus are disambiguated. Entities in 1641 are linked to DBPedia, we manually create a knowledge base from it as the coverage was not good, deduplicate entries and create new entries for instances that were linked to NIL.

Please refer to [3] for more information.

II. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

This dataset was created to evaluate entity linking performance on domains where no Wikipedia/Wikidata coverage is given and which have more diverse and difficult texts. We also want to evaluate human-in-the-loop entity linking. Existing datasets mostly are newswire and link to Wikidata or other large open-domain knowledge bases.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

This data has already been used in [3] .

C. What (other) tasks could the dataset be used for?

Entity linking, named entity recognition.

D. Who funded the creation dataset?

This work was supported by the German Research Foundation under grant No. EC 503/1-1 and GU 798/21-1 as well as by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1816B (CEDIFOR).

E. Any other comment?

No.

III. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

This dataset contains several text documents that are annotated with entity ids linking to different knowledge bases.

B. How many instances are there in total (of each type, if appropriate)?

Corpus	#Docs	#Tokens	#Entities
WWO	74	1,461,401	14,651
1641	16	11,895	480

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each data instance consists of a span that contains a named entity and a link to a knowlege base.

D. Is there a label or target associated with each instance? If so, please provide a description.

Named entities are labeld with a link linking to a knowledge base.

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no relationships.

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

It contains a subset of the Women Writers corpus that was annotated by the Women Writers project with named entities and a subset of the 1641 depositions that was annotated against DBPedia. It is a representative sample in both cases.

H. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The splits were randomly done on document level.

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

No errors that we know of.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

It relies on a knowledge base but we also provide it.

Any other comments?

No.

IV. Collection Process

A. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We did not create annotations, but just used existing datasets, converted it and manually created a knowledge base for 1641 based on the 1641 gold data.

B. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

We did not create annotations, but just used existing datasets and converted it. For 1641, we used the existing annotations and created a knowledge base from it. For entities that were linked NIL, we created a new KB entry.

C. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

We did not sample.

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Only the authors of the paper were involved.

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

We did not collect the data, we just created a new dataset from existing data.

V. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Please refer to [3] for the exact steps.

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

It is available under `https://github.com/UKPLab/acl2020-interactive-entity-linking`

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

It achieves what we wanted.

E. Any other comments

No.

VI. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

It is available under `https://github.com/UKPLab/acl2020-interactive-entity-linking` and `https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2316`

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

It is avaiable already under ASF.

C. Are there any copyrights on the data?

Yes, copyright by the 1641 depositions team and Women Writers project.

D. Are there any fees or access/export restrictions?

No.

E. Any other comments?

No.

VII. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

UKP Lab TU Darmstadt.

B. Will the dataset be updated? If so, how often and by whom?

No.

C. How will updates be communicated? (e.g., mailing list, GitHub)

No updates.

D. If the dataset becomes obsolete how will this be communicated?

Yes.

E. Is there a repository to link to any/all papers/systems that use this dataset?

No.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Yes, they can open an issue on Github.

VIII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Double blind review during a conference.

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

No.

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Not applicable.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Not applicable.

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Not applicable.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Not applicable.

I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Not applicable

J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Not applicable

K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Not applicable

M. Any other comments?

No.

REFERENCES

[1] Gary Munnelly and Seamus Lawless. Investigating Entity Linking in Early English Legal Documents. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18*, pages 59–68, 2018.

[2] John Melson and Julia Flanders. Not Just One of Your Holiday Games: Names and Name Encoding in the Women Writers Project Textbase. White paper, Women Writers Project, Brown University, 2010.

[3] Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. From zero to hero: Human-in-the-loop entity linking in low resource domains. In *The 58th annual meeting of the Association for Computational Linguistics (ACL 2020)*, April 2020.