breve and claro Datasheet

Adrian de Wynter
adewynter@microsoft.com

Anthony Hevia
anthonyhevia@microsoft.com

Si-Qing Chen
sqchen@microsoft.com

4th June 2023

Motivation

Who built them?

This was built by Adrian de Wynter, Anthony Hevia, and Si-Qing Chen at Microsoft.

Who funded this work?

Microsoft.

Extra comments

N/A

Composition

How many instances are there?

BrevE: 3, 729 sentences; (train/test) 2, 929/800.
CLaro: 3, 626 sentences; (train/test) 2, 826/800.
CLaro's final test overlap with BrevEis 13.5%.

Is the data sampled?

Yes, the data is sampled from the original corpora (OSCAR and a subset of the CWI 2018 Shared Task).

What is each instance made of?

Text and label.

What is the instance's label?

$\{0, 1\}$ for both datasets. Label=1 can be interpreted as needing simplification in its task; label=0 as not needing further simplification based on inter-annotator agreement.

Is any information missing from the instances?

No.

Is the data split?

Yes, there's validation and training (randomly-sampled) and test (cherry-picked for informativeness).

Are there any errors, sources of noise, or redundancies in the datasets?

The data is not noisy–it has been verified for grammaticality and label accuracy–but, like all corpora related to *user preference*, it is an approximation. See Preprocessing (below) and our paper for more details on this.

Are the datasets self-contained?

Yes.

Is it possible to identify people or confidential data in the datasets?

No, the data has been pseudonymised. Still, check Distribution below in case we missed anything.

Is there sensitive or harmful/biased/toxic content in the datasets?

No. The annotators flagged the sentences that were harmful, and we dropped these.

Extra comments

N/A

Collection Process

How did you collect the data?

Sampled it from OSCAR + CWI, simplified it with LLMs, and then got it annotated.

The sampling was done by collecting sentences based on specialised lists, and then randomly sampling an appropriate subset based on budget. Simplification was with GPT-3.51 and a model based on Turing. Second simplification is part of the labelling process.

Who was involved in the data collection process?

The authors collected the raw data and simplified it the first time. For the original corpora (OSCAR/CWI), the details are unknown to the authors of BrevE/CLaro.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances?

N/A, static datasets.

Were any ethical review processes conducted?

Not for BrevEand CLaro. The other datasets' ethical processes are unknown to the authors of this datasets.

Was this data collected from individuals?

No.

Has an analysis of the potential impact of the datasets and its use on data subjects (e.g., a data protection impact analysis) been conducted?

Yes, all open-sourced work by Microsoft goes through this review.

Extra comments

N/A

Preprocessing And So On

What was the preprocessing/cleaning/labeling of the data?

Yes, look at our paper for a detailed outline. In a nutshell: Annotation used a specific rubric (availabe in this repo!). Annotators were five native speakers with training in linguistics, remunerated at a rate starting at $20 USD/hr. Every annotator provided feedback on the simplifications, and suggested rewrites. We integrated these and then re-submitted for final QA and labelling. Labels are based on the inter-annotator agreement (Fleiss' $\kappa > 0.8, 0.6$) for BrevEand CLaro, respectively. The test set was built by selecting the original simplification, or the target, from the original aligned corpus when mapped to BrevE. We then randomly replaced the selection with CLaro's opposite (target to source or viceversa).

Will the raw data be included?

No, but please look at our paper to reproduce (and expand!) the datasets.

Is the software used to process the data available?

No.

Extra comments

N/A

Uses

Have the datasets been used for any tasks already?

Yes, please see our paper.

Is there a place that links to any or all papers or systems that use the datasets?

Not beyond this one.

Is there anything about the composition of the datasets or the way it was collected and preprocessed/cleaned/labeled that might impact future use?

No.

Are there tasks for which the datasets should not be used?

Do not use this datasets to infringe in other people's liberties–this includes surveillance, tracking, etcetera.

Extra comments

N/A

Distribution

Will the datasets be distributed to third parties? How and when will it be distributed?

Yes, the datasets is freely available in this repository, starting June 5th, 2023.

What licence are these datasets under?

Please see the licence file in this repository.

Have any third parties imposed IP-based or other restrictions on the data?

To our knowledge, no. But we support takedown requests. Please open an issue in this repository if you need it.

Do any export controls or other regulatory restrictions apply to the datasets or to individual instances?

No.

Extra comments

N/A

Maintenance

Whom will be maintaining the datasets? How can they be contacted?

Please open an issue in this repository. The appropriate owner will be contacted based on your request.

Are there any errata?

No. We will update this section if applicable.

Will the datasets be updated?

Takedown requests aside, there will not be any updates. If there are any, we will communicate them in the README file.

If the datasets relates to people, are there applicable limits on the retention of the data?

N/A

Will older versions of the datasets continue to be supported?

No, but they will be accessible via Github's version control.

How can people build on or contribute to the datasets?

The datasets are released as CC-4.0. You can reuse them subject to the licence terms. We also welcome contributions via this repository.

Extra comments

N/A

Adrian de Wynter, Microsoft. Last updated: 4th June 2023.