News-Commentary v15 Dataset Datasheet

A collection of parallel corpora

I. Disclaimer

First of all, notice that this Datasheet[1] is not filled out by the dataset author/creator. Therefore it is strongly recommended to only make use of this if the author/creator has not filled in a proper Datasheet or to use it in combination.
One may found all the details about the writer(s) of this Datasheet hereunder. Furthermore, one may found the different sources used by the writer, from which information was retrieved, referenced all throughout the document. And of course, those questions for which there is not validated information available will be left empty.

This Datasheet has been filled out by Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia and Margarita Geleta, a group of junior year Data Science and Engineering undergraduate degree students at UPC in Barcelona.

II. Motivation For Datasheet Creation

A. Who created the dataset(e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The parallel corpus of News Commentaries was provided by the Association for Computational Linguistics (ACL).

B. Did they fund it themselves? If there is an associated grant, please provide the name of the grant or/and the grant name and number)?

The ACL covered the expenses, however, there is no information whether there was an associated grant or not.

C. For what purpose was the dataset created? Was there a specific task in mind? If so, please specify the result type (e.g. unit) to be expected.

Different versions have been released since the first time this dataset was published. The original purpose was for the Conference for Statistical Machine Translation Evaluation Campaign. Then, new versions have been released aiming to update the dataset by means of crawling the Project Syndicate website[2].

The main task from which this dataset was intended to be used were Machine Translation (MT) and Evaluation.

D. Could any of these uses, or their results, interfere with human will or communicate a false reality?

It is known that the results in MT task can potentially communicate biased or unfair realities[3].

The dataset is composed of political and economic commentaries from the Project Syndicate website, which is considered to be *The World's Opinion Page*. Thus, the source can inherit bias or unfairness from these articles related to the political and economic condition over the globe.

E. What is the antiquity of the file? Provide, please, the current date.

News Commentary Parallel Corpus was first released in 2012. The released date of the latest versions, 15th version, dates in 2020[4].

F. Has there been any monetary profit from the creation of this dataset?

The dataset was released aiming to be useful for the Computational Linguistic research community in the field of Natural Language Processing (NLP). Nevertheless, information whether explicit monetary profit has been made or not could not be found.

III. Datasheet Composition

A. Are there multiple types of instances or is there just one type? Please specify the type(s) (e.g. Raw data, pre-processed, continuous, discrete)

All information found in the dataset consists of pieces of tokenized and untokenized (both options can be downloaded) **plain text** for both parallel and monolingual corpora.

[1]This Datasheet has been inspired by Datasheets for datasets[1], including questions that could be answer by the writer. On top of that, it also includes specific questions that can apply to corpora used in Machine Translation.

[2]https://www.project-syndicate.org

[3]https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html

[4]http://data.statmt.org/news-commentary/

B. What do the instances (of each type, if appropriate) that comprise the dataset represent (e.g., documents, photos, people, countries)?

They represent separate sentences and words. In the parallel corpus rows represent pairs of these mentioned types that have been aligned between the 2 languages.

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

The number of instances in parallel corpora highly depends on the languages that compose it. The range of instances varies from 161k (Indonesian - Kazakh) to 47M (Arabic - English)[5]. Furthermore, there are some language pairs that have no references translated.

For the monolingual corpora, taking into account that not all languages have one, the range of instances varies from 205k for Japanese to 36M for English[6].

D. Does the dataset contain all possible instances or is it just a sample of a larger set?

It contains political and economic commentary[7] crawled from the already mentioned Project Syndicate website. It is not a subset of any other published dataset, but there exist several versions which are based on different crawling executions.

E. Is there a label or a target associated with each of the instances? If so, please provide a description.

No, because it is just a recompilation of plain text. Sentences can be detected and after a simple pre-processing pass that detects punctuation (for example), but no explicit labels exist.

F. What is the format of the data (e.g. .json, .xml, .csv)?

Files used in the monolingual corpus consist of .txt sources of plain text. The alignment units are saved in TMX files or also (more interpretable) in XML files, where the alignment between the positions of words in the sentences pairs is saved.

G. 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.

Information provided in the dataset only covers data that has appeared in the Project Syndicate website.

Also, there are language pairs for which its parallel corpora is totally empty, for example: Indonesian - Japanese[8].

H. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Do not include missing information here.

There might be sentences that appear more than once in different news and contexts, although it is not considered a redundancy because it might be useful for some models or problems to understand the different applications of a specific sentence. With low probability, it could happen that a sentence in a different language from the one used in a given dataset might appear.

I. Is there any verification that guarantees there is not institutionalization of unfair biases? Both regarding the dataset itself and the potential algorithms that could use it.

No, there is no verification. All the bias that is contained in the compilation of text from political and economic news will be present in the algorithms built on top of it.

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

For MT, the standard partition is useful. Small subsets of the parallel corpus are created for both validation and testing. These allows the model to test if it is able to perform a phrase to phrase translation without knowing the reference alignment.

K. 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 consists of political and economic news crawled from *Project Syndicate*. Once these are published they might never be modified again (maybe in case there was an error, which sounds unlikely since these are carefully reviewed

[5] http://data.statmt.org/news-commentary/v15/training/
[6] http://data.statmt.org/news-commentary/v15/training-monolingual/
[7] http://www.casmacat.eu/corpus/news-commentary.html
[8] http://data.statmt.org/news-commentary/v15/training/?C=S;O=A

before being published). There is no fee since it is and open source of news.

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

No it doesn't. As mentioned before they come from an open source of information.

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

It should not be the case, since the exact same information is shown to the public that reads the news through this source.

N. Does the dataset relate to people? If so, please specify a) Whether the dataset identifies sub-populations or not. b) Whether the dataset identifies individual people or not. c) Whether it contains information that could vulnerate any individuals or their rights. c) Any other verified information you can provide.

It might do but not in a targeted way, the people or individual affected in every single article depends on the topic of it and the relevant cause that has made it appear publicly.

O. Does the dataset cover included languages equitatively?

Since they are splitted according to language pairs, every one of these must have the same sentences in the parallel corpus, so yes. Another thing is that an extra monolingual corpus of a single language of the pair is added, but it is done for the purpose of translation.

Notice that the size of crawled data and the topics covered in the articles depend on the different languages that are used. Hence, they do not output exact copies of the same dataset nor same amount of data.

P. Is there any evidence that the data may be somehow biased? (e.g. towards gender, ethics)

It is widely criticised that newspapers give a subjective point of view in every new they publish, so there's the awareness that the possible bias in Project Syndicate relating politics and economics might be present.

Q. Is the data made up of formal text, informal text or both equitably?

It is all formal text. It is written in the manner one would like to see in an online newspaper.

R. Does the data contain incorrect language expressions on purpose? Does it also contain slang terms? If that's the case, please provide which instances of the data correspond to these texts.

As mentioned before, it should not take into account the source where the text comes from.

IV. Collection Process

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

It all comes from the crawling process of the same source (website).

B. Are there any guarantees that prove that the acquisition of the data did not vulnerate any law or anyone's rights?

The fact that all this data could be accessed at any given moment in the website it is extracted from implies that the acquisition if legal. If not used violating any of the website terms and conditions, then the whole process is also legal.

C. Are there any guarantees that prove that the data is reliable?

According to the size of the dataset, it can be said that it can consistently represent a language, so a language model can be built with this data and no more.

D. Please, if the dataset relates to people, specify any information regarding these people. (e.g. Was the data collected from these people directly? Did all the involved parts give their explicit consent? Is there any mechanism available to revoke this consent in the future, if desired?)

There is the same security that exists in the newspaper polices. Any reclamation that might be sent to the newspaper after the crawling is done might affect the ethical purity of the data collection.

E. 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.

No, no analysis regarding this subject has been conducted.

F. Were any ethical review processes conducted?

No, there were not any ethical review processes conducted.

G. Please specify any other information regarding the collection process (e.g. Who collected the data, whether they were compensated, what mechanisms were used) Please, only include if verified.

The dataset release purpose was for the Conference for Statistical Machine Translation Evaluation Campaign. No information about a possible compensation is available - even though it seems improbable. The mechanism consisted of a crawling process on the Project Syndicate website.

H. Does the data come from a single source or is it the result of a combination of data coming from different sources? In any case, please provide references.

The data comes from a single source, which is the Project Syndicate website, as stated here[9].

I. If the same content was to be extracted from a different source, would it be similar?

Yes, the data would probably be very similar - given that newspapers' content tends to be about similar topics and thus, the used words are very similar. If the new source were to be different from a newspaper, there could be some significant differences when it comes to words use distribution and some specific translations. Even so, the law of large numbers implies that, if the source contains a high enough amount of data, the words distribution will tend to be very coincident between sources.

V. Data Preprocessing

A. Please specify any information regarding the preprocessing that you may know (e.g. the person who created the dataset has somehow explained it) or be able to find (e.g. there exists and informational site about the dataset). Please, only include if verified. (e.g. Was there any mechanism applied to obtain a neutral language?, Were all instances preprocessed the same way?

The corpus is available in several formats: raw text files (not preprocessed) and sentence aligned files (with alignment preprocessing). No further information about preprocessing is given by the dataset creators.

B. Was there any mechanism applied to obtain a neutral language?

This is no specified by the dataset creators.

C. Were all instances preprocessed the same way?

The entire corpus has been preprocessed the same way.

VI. Dataset Uses

A. Has the dataset been used already? If so, please provide a description.

The corpus has been used as training data for the CASMACAT project (2012-2014) which built a translator's workbench to improve productivity, quality, and work practices in the translation industry. CASMACAT stands for *Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation*.

Also, this dataset is used in many WMT conferences and workshops, which are annual events on machine translation.

B. Is there repository that links to any or all papers or Systems that use dataset? If so, please provide a link or other access point.

No, there is not. The official site of the dataset does not link to any repositories nor works using this dataset.

C. What (other) tasks could the dataset be used for? Please include your own intentions, if any.

The dataset is intended for machine translation, even though it can also be used for other NLP tasks such as language modelling or language generation.

D. Are there tasks for which the dataset should not be used? If so, please give a description.

There are no explicit tasks where the dataset should not be used, since it can be used to do fine-tuning or domain adaptation in other tasks.

E. Any other comments? (e.g. Do the collection or preprocessing processes impact future uses?)

There is minimal risk for harm: the data was already public.

VII. Dataset Distribution

[9]http://www.casmacat.eu/corpus/news-commentary.html

A. Please specify the source where you got the dataset from.

The source of the dataset are all the pages contained within the Project Syndicate[10] website.

B. When was the dataset first released?

The release (the upload) of this dataset version (v15) was on 10th January, 2020.

C. Was the dataset released publicly? Are there any regions where this dataset is not available?

Yes, the dataset was released publicly. No, the dataset is available for all regions with access to the Internet.

D. Is the dataset distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? Please cite a verified source.

As stated by the authors, "No claims of intellectual property are made on the work of preparation of the corpus". See source.

E. Any other comments? (e.g. How has the data been distributed? Who has access to the dataset? When was the dataset first distributed? Are there any other regulations on the dataset?)

The distribution of the first version of the dataset was on year 2012.

VIII. Dataset Maintenance

A. s there any verified manner of contacting the creator of the dataset?

There is no verified manner of contacting the creator as the latter is not known. The maintainer and responsible of the entire OPUS corpus can be contacted at jorg.tiedemann@helsinki.fi.

B. Has there any erratum been notified?

No erratum has been notified. Data itself can not contain errors, as it is simply the results of crawling an existing web, and it can be accessed with no errors by means of the provided files.

C. Is there any verified information on whether the dataset will be updated in any form in the future?

There is no verified information on this matter. Even so, the latest updates have been recent which might indicate there is still room for more.

[10]https://www.project-syndicate.org

D. Could changes to current legislation end the right-of-use of the dataset?

There are low risks in this direction, but some of them might depend on the legislation of web crawling which can limit the methodology used for the data extraction, as well as restricting the possible uses of it (specially related to obtaining profit from the data).

E. Is there any mechanism, that promotes vocabulary enrichment, automatically developed?

No, there is no such mechanism. One of the main reasons being the fact that the dataset is a reflection of a website, and thus it simply contains what there is in it - with no external additions nor enrichment.

F. Any other comments? (e.g. Will the dataset be updated in the future? Is there someone supporting/hosting/maintaining the dataset? If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?)

The dataset is hosted by jorg.tiedemann@helsinki.fi. It does not appear to be actively maintained. It does not relate to specific people in a direct way.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. 2020.