News-Commentary v15 Dataset Datasheet

A collection of parallel corpora

Disclaimer

This Datasheet has been inspired by [1] and modified as proposed by [2] and it is not filled out by the dataset creator. Therefore it is strongly recommended to only make use of this if the creator has not filled in a proper datasheet or to use it in combination. It is required that writers indicate their personal and contact data as well as the date this datasheet was last reviewed hereunder. Please, also remember to change the datasheet title to the name of the datatset in question.

This datasheet has been filled out by Marta R. Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia and Margarita Geleta, from Universitat Politècnica de Catalunya (UPC) in Barcelona, which can be contact at `marta.ruiz@upc.edu`.

This datasheet was last reviewed on May 25th 2020. Authors want to specially thank Barry Haddow for his feedback on this datasheet.

I. Motivation

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 University of Edinburgh, specifically the machine translation group [1].

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

The University of Edinburgh covered the expenses. It has never been attached to a specific grant.

C. For what purpose was the data set 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 WMT International Evaluation Campaign [2]. Then, new versions have been released aiming to update the dataset by means of crawling the Project Syndicate website[3]. The main task from which this dataset was intended to be used was Machine Translation (MT).

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], [4]. 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.

The released date of the latest versions, the 15th version, was in 2020[4]. At May 25th 2020.

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 MT but also in Natural Language Processing (NLP) in general. Nevertheless, information whether explicit monetary profit has been made or not could not be found.

II. Composition

A. Is there any synthetic data in the dataset? If so, in what percentage?

Since this dataset is defined as a vanilla compilation of text, it does not contain synthetic data.

B. Are there multiple types of instances or is there just one type? Please specify the type(s), e.g. Raw data, preprocessed, symbolic.

All information found in the dataset consists of pieces of plain text for both parallel and monolingual corpora.

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

[1] http://www.statmt.org
[2] http://www.statmt.org/wmt20/
[3] https://www.project-syndicate.org
[4] http://data.statmt.org/news-commentary/

They represent sentences which, in some cases, are compound of a single word. In the parallel corpus rows represent pairs of these mentioned types that have been aligned between two languages.

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

The number of instances, sentences, in parallel corpora highly depends on the languages that compose it. The range of instances varies from around 1,500 sentences (Japanese - French) to almost 400,000 sentences (Spanish - English)[5]. As for this source, there is parallel corpus among 12 languages: Arabic, Czech, German, English, Spanish, French, Italian, Japanese, Dutch, Portuguese, Russian and Chinese.

For the monolingual corpora available, the range of instances varies from few thousands for Japanese to more than 600 thousand sentences for English[6].

E. Is the dataset self-contained? i.e. Is the dataset different than an original one due to the preprocessing process? In case this dataset is a subset of another one, is the original dataset available?

It contains political and economic commentaries[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. Therefore, although the dataset is not a subset of any other, it is a reduced crawling from the Project Syndicate website that only takes in to account the topics of Economics and Politics.

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

Only information at the document level exists, except for that it is just a recompilation of plain text.

G. 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.

H. 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.

There are some language pairs for which there is no parallel corpus at all, for example the Indonesian - Japanese[8].

[5]http://opus.nlpl.eu/

[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

I. 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 can happen that in a given dataset may exist an instance, a sentence, in a different language than the expected.

J. 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.

K. 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.

There are development and test sets prepared for each WMT International Evaluation Campaign [5].

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

M. 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 does not. As mentioned before they come from an open source of information.

N. 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.

O. Does the dataset relate to people? If so, please specify a) Whether the dataset identifies subpopulations or not. b) Whether the dataset identifies indivual people or not. c) Whether it contains information that could vulnerate any individuals or their rights. c) Any other verified information on the topic that can be provided.

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

P. Does the dataset cover included languages equally?

It is obvious that in parallel corpora both of the languages in question are equally covered.

However, as we have stated before, not all parallel corpora have the same size. As well as, if we compare monolingual corpora, the included languages are not covered equally. This is due to the lack of data in such a language in the crawling performed and therefore there is not a deliberate misbehaviour in any action regarding the dataset creation.

Because of this same reason, one should also notice that the content found in the different parallel corpora or monolingual corpora may not be the same.

Q. Is there any evidence that the data may be somehow biased? i.e. towards gender, ethics, beliefs.

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

R. 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 expect to see in an online newspaper.

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

As mentioned before, it should contain appropriate language expressions due to the fact that is crawled text from an online newspaper.

III. Collection Process

A. Where was the data collected at? Please include as much detail; i.e. country, city, community, entity and so on.

The text instances in dataset were crawled from the Project Syndicate website[9] which is compound of articles that talk about the economical, political, developing, sustainability-related, cultural or innovation situations of places around the world and, of course, these articles are written by many journalists, of all nationalities. Therefore the data may be considered to have been extracted from many places around the globe.

B. If the dataset is a sample from a larger set, what was the sampling strategy? i.e. deterministic, probabilistic with specific sampling probabilities.

It all comes from the crawling process of the same source.

C. Are there particular measures taken to avoid law violation or violation of 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 is legal. If not used violating any of the website terms and conditions, then the whole process is also legal.

D. Are there any particular measures taken to make the data reliable?

Despite being no guarantees, the fact that the data source is the Project Syndicate makes the data pretty reliable.

E. Did the collection process involve the participation of individual people? If so, please report any information available regarding the following questions: Was the data collected from people directly? Did all the involved parties 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.

F. Has an analysis of the potential impact of the dataset and its use on data subjects been conducted? i.e. a data protection impact analysis. 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.

G. Were any ethical review processes conducted?

No, there were not any ethical review processes conducted.

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[10] however, the dataset was created by researchers at the University of Edinburgh.

I. Can this data be extracted independently?

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

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

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.

J. Please specify any other information regarding the collection process. i.e. Who collected the data, whether they were compensated or not, what mechanisms were used. Please, only include if verified.

The dataset release purpose was for the WMT International Evaluation Campaing. Nothing suggests that there has been any economical compensation. The mechanism consisted of a crawling process on the Project Syndicate website.

IV. Preprocessing/Cleaning/Labelling

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). Please, only include if verified. i.e. 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.

The entire corpus has been preprocessed the same way.

V. Uses

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

This dataset has been used for all the editions of the WMT International Evaluation Campaings [5], which are annual events on MT. This is its main use, although it has also been used in many research papers.

The dataset has also been used in several projects, for example, 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* [11].

[11] http://www.casmacat.eu/

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

The WMT event presents a findings paper for each edition [5], but beyond this, there is no specific repository that contains information of this kind.

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

The dataset is intended for MT, even though it can also be used for other NLP tasks such as language modelling or language generation. We are only completing this datasheet for the sake of providing it to the community. But we are not going to actually use the dataset for any task.

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

There are no explicit tasks where the dataset should not be used.

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

There is minimal risk for harm since it is publicly available data.

VI. Distribution

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[12] website.

B. When was the dataset first released?

The initial release of this corpus was back in 2007[13].

C. Are there any restrictions regarding the distribution and/or usage of this data in any particular geographic regions?

The dataset was released publicly and therefore it is available for all regions.

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.

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

[13] http://www.statmt.org/wmt07/shared-task.html

VII. Maintenance

A. Is 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 corpus can be contacting Barry Haddow (University of Edinburgh) at bhaddow@inf.ed.ac.uk.

B. Specify any limitations there might be to contributing to the dataset. i.e. Can anyone contribute to it? Can someone do it at all?

There are not specific limitations, any contribution can be done by directly contacting Barry Haddow (as mentioned in the previous question). Mainly for notifying on errors.

C. Has any erratum been notified?

No erratum has been notified. Data itself cannot 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.

D. Is there any verified information on whether the dataset will be updated in any form in the future? Is someone in charge of checking if any of the data has become irrelevant throughout time? If so, will it be removed or labeled somehow?

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

E. Is there any available log about the changes performed previously in the dataset?

There is not any log of this type, in words of one of its creators (Barry Haddow), this is because the main changes are about enlarging the data.

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

In principle, there are low risks in this direction. But we could consider changes in the legislation of web crawling, since this could limit the methodology used for the data extraction. As well as changes in the restrictions of usage, specially the ones related to obtaining profit from the data, which could invalidate some tasks performed on it.

G. Are there any lifelong learning updates, such as 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.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for Datasets". 2018.

[2] Marta R. Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia, and Margarita Geleta. MT-Adapted Datasheets for Datasets: Template and Repository, arXiv: 2005.13156. May 2020.

[3] Marcelo O.R. Prates, Pedro H. Avelar, and Luís C. Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Comput Applic*, page 6363–6381, 2020.

[4] Joel Escudé Font and Marta R. Costa-jussà. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy, August 2019. Association for Computational Linguistics.

[5] Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics.