Europarlv10 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(s), 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 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 Europarl parallel corpus was created by a group of researchers led by Philipp Koehn at the University of Edinburgh[2].

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

The construction of the European Parliament Proceedings Parallel corpus was supported by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme).

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.

Parallel Corpus was created to provide a training data for statistical machine translation (SMT). However, nowadays there is barely no research on SMT, since almost everything is done with Neural Machine Trasnlation (NMT). And therefore we should be able to assume that this corpus is suggested to be used for these NMT tasks. Consequently the expected result would be the translation of the input, and therefore, text data.

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

Although there has not been made any specific communication on this topic, from our background knowledge we can be sure to warn that, due to the existence of bias in MT, the output text could be affected in such a way that it would not be portraying the actual essence of the translation.

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

The initial release of this corpus was back in 2005 and consisted of data up to 2001. It has since then been updated many times. The last updated was on January the 17th 2020 and the data contained goes up to November 2011.

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

This data was collected mainly to aid the authors' research in SMT and there is not evidence of getting any other profit from it.

G. Any other comments?

The authors have shared that they "used the corpus to build 110 machine translation systems for all the possible language pairs. The resulting systems and their performances demonstrate the different challenges for statistical machine translation for different language pairs"[3].

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, preprocessed, continuous, discrete)

At first the corpus was compound of raw data that had been crawled from the available content in the web. However, after the preprocessing process this converted in to aligned sentences. Moreover, the dataset also contains some metadata like the source, target, file ID, chapter ID, speaker ID, speaker name, language, and affiliation. Which are strings and integers.

[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.

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

Each of the columns indicates its content by the name (i.e. chapter ID indicates what document refers to) and then the corpus itself contains the corresponding text data in several source and target languages.

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

The whole dataset, after sentence aligning and removing XML, consists of 24090883 aligned sentences.

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

Although there may be more records on the proceedings of the European Parliament than the ones included in the corpus. It is self-contained in the sense that it is compound of all data that the creators retrieved to create the Europarl corpus.

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

Yes, in this dataset there is a target language that is the language being translated to. The source and target languages are diverse, with up to 20 different languages.

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

The data comes in a TSV (tab-separated values) file.

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.

It has not been reported any issue of this kind.

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.

It has been noticed that some special HTML entities and noisy characters have not been removed from the whole set of the Europarl corpus[2].

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.

Since political discourses may portray aspects like personal opinions or generalizations, these could, either intentionally or not, generate biases on the data that could perpetuate throughout the translations. No mechanism to avoid biases has been used in the preprocessing process. Therefore, both because it is data from politics and because it is aimed to be used in MT could present bias in its content and also in the output it may produce.

J. 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.

It is recommended to use data from the last quarter of 2000 as a test set, while the rest should be used as training data[2].

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.

The dataset is self-contained in terms previously explained and therefore it is linked to the website of the European Parliament[4]. As it is compound of past records, there is an implicit guarantee that the data will remain constant. There are official reportings that can be found in the website of the European Parliament[4]. The authors assure that they are not aware of the existence of any copyright restrictions of the material but encourage those that use the corpus to contact Philipp Koehn at pkoehn@inf.ed.ac.uk as indicated in the source website[2].

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

As data comes from European Parliament speeches, which are all available to the public[4], no confidential data is present.

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 does not. However, several political points of views are present in the data, which may not be on par to someone's political views or may be considered extreme.

N. 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 you can provide.

Because of the nature of the data, the corpus may reference many different people. On the other side, the data contained

in the dataset, not only the corpus, does reference the individual speakers. However, it does not identify subpopulations and neither could vulnerate anyone's rights.

O. Does the dataset cover included languages equally?

Although 20 languages are covered by this dataset, they are not covered equally. As an example, while the FR-EN corpus contains about $2 \cdot 10^6$ instances, the CZ-EN corpus contains about $4 \cdot 10^5$ observations.

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

Apparently, there is not, but as stated earlier, political-induced biases may be present.

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

Mostly formal since it is compound of speeches coming from Proceedings at the European Parliament.

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.

One can assume it does not in the lines to which political speeches tend to be.

S. Any other comments?

There are not any other specifications to be made regarding the dataset content.

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)

No specific sampling was performed as the dataset is not part from a larger set, as previously explained.

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

As the source data has been made public by the European Parliament itself, one should be able to assume that it is all in compliance with European laws.

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

Despite being no guarantees, the fact that the data source is the European Parliament itself makes the data pretty reliable.

D. Does the dataset realte to people? 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?)

It does relate to people as every observation includes who is the speaker. The speaker name along with its affiliation is provided. The participants concern the gathering of the data according to European laws, as they are participating in public acts. As part of the public work of a civil servant, any potential mechanism that could exist for the speaker to revocate the data provided would be under the European Parliament's concern.

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.

We have not found evidence of any analysis in this direction.

F. Were any ethical review processes conducted?

No ethical review processes were conducted, apparently.

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 acquisition of this parallel corpus for use in a statistical machine translation system typically takes five steps[3]:
- Obtain the raw data (e.g., by crawling the web)
- Extract and map parallel chunks of text (document alignment)
- Break the text into sentences (sentence splitting)
- Prepare the corpus for SMT systems (normalisation, tokenisation)
- Map sentences in one language sentences in the other language (sentence alignment)

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 that makes up the corpus was extracted from the website of the European Parliament[4] and then prepared for linguistic research. After sentence splitting and tokenization the sentences were aligned across languages with the help of an algorithm developed by Gale & Church[5].

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

It definitely should, since it is official public data.

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?

Matching items were extracted and labeled with their corresponding document IDs. Using a preprocessor [3] sentence boundaries were identified. The data was sentence-aligned using a tool based on the *Church and Gale algorithm*.

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

This is not specified by the corpus creators and therefore we assume it was not part of the preprocessing process.

C. Were all instances preprocessed the same way?

The whole corpus has been preprocessed the same way.

VI. Dataset Uses

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

The dataset is free and it is available for commercial use and for research purposes. Surprisingly, bibliometric analyses show that it has hardly been used in translation studies, although this was the first purpose of its creators. Toolkits such as *EuroparlExtract* have been developed with this dataset [6]. As can be found in the paper Europarl: A Parallel Corpus for Statistical Machine Translation[3] "it has been used for many other natural language problems: (such as) word sense disambiguation, anaphora resolution, information extraction, etc."

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

No, there is not. The source site of the dataset does not link to any repositories nor works using this dataset and the authors have not referenced any either[3].

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

The dataset is intended for Machine Translation tasks. Yet the original paper[3] puts emphasis on SMT (Statistical Machine Translation) tasks, this dataset could also be used for NMT (Neural Machine Translation), NLP (Neural Language Processing) or any other language modelling or language generation tasks.

D. Are there tasks for which the data set 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 since the data is already public under the reponsibility of the European Parliament.

VII. Dataset Distribution

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

We have retrieved the corpus from the "Fifth Conference on Machine Translation WMT20" website [7].

B. When was the dataset first released?

As stated before, the initial release of this corpus was back in 2005.

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

Yes, the dataset was released publicly. It is freely available on the European Parliament website[4], and therefore it should be available in 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.

Stated by the authors: "We are not aware of any copyright restrictions of the material"[7].

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

No additional comments can be supplied.

VIII. Dataset Maintenance

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

Yes. All questions and comments can be sent to Philipp Koehn at pkoehn@inf.ed.ac.uk

B. Has there any erratum been notified?

No specification has been made on the existence of errata in the Europarl corpus.

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

Despite not being ensured, the wide use of the dataset and the fact that it has been updated more or less in a yearly basis during the last years, makes it highly probable that this data will be updated in the future.

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

Not in the foreseeable future, as it would imply that the European Parliament's data would not be public anymore.

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

Not that we are aware of, as the preprocessing steps only include, mainly, tokenization and alignment[3].

F. Any other comments? (e.g. 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?)

No additional comments can be supplied.

REFERENCES

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

[2] Title:"Europarl corpus for WMT20".

[3] Philipp Koehn. Title:"Europarl: A Parallel Corpus for Statistical Machine Translation". 2005.

[4] Title:"European Parliament Website".

[5] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102, March 1993.

[6] Michael Ustaszewski. Title:"Optimizing the Europarl corpus for translation studies with the EuroparlExtract toolkit". *Perspectives*, 27(1):107–123, 2019.

[7] Title:"Fifth Conference on Machine Translation WMT20".