Europarl v10 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 contacted 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 Europarl parallel corpus was created by a group of researchers led by Philipp Koehn at the University of Edinburgh [3].

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 construction of the European Parliament Proceedings Parallel corpus was supported by the EuroMatrix project and later by the following EuroMatrixPlus project funded by the European Commission (7th Framework Programme). It was also funded by the Mosescore[1] project which "encourages the development and usage of open source machine translation tools" and which is also supported by the European Commission Grant Number 288487 under the 7th Framework Programme.

[1]Find more about the Mosescore project at http://www.statmt.org/mosescore/ and https://github.com/moses-smt/mosesdecoder

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.

EuroParl Parallel Corpus was created to provide a training data for corpus-based Machine Translation (MT) systems such as 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, interfere with human will or communicate a false reality?

Although we are not aware of 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 [4], the output text could be affected in such a way that it would not be portraying the actual essence of the translation. For example in the case of translating a sentence for which there is not explicit information on the subject's gender in one language while the translation must contain this gender information, due to the way gender is expressed in each of the languages[5].

Aside from this, one should also notice that the errors encountered in the dataset (e.g. misalignments, mixing of languages) could also report misleading results.

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 update was on January the 17th 2020 and the data contained goes up to November 2011. At May 25th 2020.

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

This data was collected mainly to aid the authors' research in statisitcal MT 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 MT for different language pairs"[3].

II. Composition

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

No, there is not presence of synthetic data in this dataset.

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

At first the corpus was compound of raw data that had been crawled from the available content in the web [2]. However, after the preprocessing process this is converted to aligned sentences. Moreover, the dataset also contains metadata (i.e. the source, target, file ID, chapter ID, speaker ID, speaker name, language, and affiliation, which are strings and integers).

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

The text corpus is organised in documents, each entry indicates its content by the name (i.e. chapter ID indicates what document refers to) and then the corresponding text data in several source and target 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 over few hundred thousands (Italian - Romanian) up to two million (French - English). As for this source [6], there are 21 languages available: Bulgarian, Czech, Danish, German, Lithuanian, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovenian, Swedish.

In addition to the parallel corpus, there is also monolingual corpora available, and the range of instances varies from less than half a million for both Bulgarian and Romanian up to two and a half million for English.

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?

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.

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

Yes, in addition to the plain text in two languages, which is the core content of the dataset, there is some associated metada which includes the information of the structure of the proceedings and about the speaker [7].

G. What is the format of the data? e.g. .json, .xml, .csv .

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

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.

It has not been reported any issue of this kind.

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.

It has been noticed that some special HTML entities and noisy characters have not been removed from the whole set of the Europarl corpus[3].

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.

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.

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.

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]http://www.europarl.eu.int/

[3]http://www.statmt.org/europarl/

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.

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. 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[5].

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.

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

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 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.

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.

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.

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 an example, while the French-English (FR-EN) corpus has over $1 \cdot 10^6$ instances (sentences), the Greek-English (EL-EN) corpus has a little less than $4.5 \cdot 10^5$ instances. 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.

As stated earlier, political-induced biases may be present. Furthermore, only a 30% of the sentences are utterated by women, so there is clearly a gender bias present [8] . Aside from this inherited potential bias the errors found in the dataset could also cause biased results.

R. 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.

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.

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

III. Collection Process

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

As said before, the data contains speeches that took place at the European Parliament and therefore this data was generated either at Strasbourg, France or at Brussels, Belgium. The collection of the data, however, was originally performed at the University of Edinburgh [3].

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

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

C. Are there particular measures taken to avoid law violation or violation of anyone's rights?

[4]http://www.europarl.eu.int/
[5]http://www.statmt.org/europarl/

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.

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

Despite being no guarantees, the fact that the data source is the European Parliament itself 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?

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.

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.

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

G. Were any ethical review processes conducted?

No ethical review processes were conducted, apparently.

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 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[9].

I. Can this data be extracted independently?

It definitely should, since it is official public data.

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.

As stated in [3], the acquisition of a parallel corpus for the use in MT typically takes the following 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)
- Map sentences in one language sentences in the other language (sentence alignment)
- Prepare the corpus for statistical MT systems (normalisation, tokenisation)

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?

As mentioned, matching items were extracted and labeled with their corresponding document IDs [3]. As mentioned, sentence boundaries were automatically detected by using a preprocessing sentence boundaries. The data was sentence-aligned using the Gale & Church proposed algorithm [9].

The whole corpus has been preprocessed the same way.

V. 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 [10].

Some works using this dataset include: "*Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish*" by Adrià de Gispert and José B. Mariño, and "*Bilingual Sentence Alignment of a Parallel Corpus by Using English as a Pivot Language*" by Josafá de Jesus Aguiar Pontes.

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.

There is not such specific repository. However, one can find the annual proceedings of WMT [6], with many papers reporting results based on this dataset.

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

The dataset is intended for MT tasks. The original paper [3] puts emphasis on statistical MT tasks, but more

[6]http://www.statmt.org/wmt19/papers.html

generally this dataset is widely used for corpus-based MT, and nowadays is mostly used for Neural MT. Other NLP (Natural Language Processing) tasks such as language modeling, coreference resolution, name entity recognition can also benefit from this dataset. 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 the data is already public under the reponsibility of the European Parliament.

VI. Distribution

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

We have retrieved the corpus from the following sources: WMT [11] and Opus [6].

B. When was the dataset first released?

The initial release of this corpus was back in 2005.

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

The dataset was released publicly. It is freely available on the European Parliament website, 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 WMT source: "We are not aware of any copyright restrictions of the material.

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

A number of researchers have adapted the Europarl corpus for specific purposes in order to further enhance its usefulness and to compensate for some of its drawbacks. For example, the organizers of WMT have been updating, removing irrelevant data and documenting all these changes from the dataset.

C. Has any erratum been notified?

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

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?

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. Although there is not a person in charge of ensuring the relevance of the data, researchers, usually, modify the dataset in order to make it more useful and meaningful for their projects.

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

Yes, all the updates can be found at the source page: Scrolling down through the page one can find the versions http://www.statmt.org/europarl/archives.html#v1log, although not every version is fully available.

F. 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.

G. Are there any lifelong learning updates, such as vocabulary enrichment, automatically developed?

Not that we are aware of.

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] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

[4] 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.

[5] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[6] Jörg Tiedemann. Opus – parallel corpora for everyone. *Baltic Journal of Modern Computing*, page 384, 2016. Special Issue: Proceedings of the 19th Annual Conference of the European Association of Machine Translation (EAMT) – Projects/Products Volume: Proceeding volume:.

[7] Alina Karakanta, Mihaela Vela, and Elke Teich. Preserving and extending metadata in parliamentary debates. In *Proceedings of the LREC*, Miyazaki, Japan, 2018.

[8] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008. Association for Computational Linguistics, 2018.

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

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

[11] 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.