Fairness in multilingual NLP

A Data Management Plan created using DMPonline.be

Creator: Miryam de Lhoneux

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Grant number / URL: C14/23/096

ID: 205570

Start date: 01-10-2023

End date: 30-09-2027

Project abstract:

Multilingual NLP is booming. This is due in no small part to language models trained on large amounts of multilingual data (such as mBERT and XLM-R) which have been found to have surprising cross-lingual transfer capabilities, in spite of receiving no cross-lingual supervision (the models are not told anything about how to transfer knowledge between languages). Despite this progress, these models only cover a fraction of the world's languages, with large inequalities in performance. The project aims to address this inequality. We view it as an issue in fairness where different languages are not treated equally. We look at the field of fairness in AI to identify the biases that are responsible for this inequality. We develop methods to address these biases combining ideas from algorithmic fairness, neuro-symbolic AI and computer vision in order to make NLP fairer with respect to typological diversity.

Last modified: 20-03-2024

Fairness in multilingual NLP

Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		*Indicate: **N**(ew data) or **E**(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
CELEX	Morphological decomposition of lemmata in English, German and Dutch.	E	D	T	csv	<1GB	
OSCAR	Scraped web text in many languages.	E	D	T	HuggingFace Dataloader (JSON internally)	> 5 TB	
Grambank	Typological database	E	D	T	CSV	< 1GB	
URIEL	Typological database	E	D	T	CSV	< 1GB	
WALS	Typological database	E	D	T	CSV	< 1GB	
Universal Dependencies	Treebanks	E	D	T	CoNLL	< 100GB	
Parallel Bible Corpus	Parallel texts	E	D	T	TSV	< 1GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

- http://celex.mpi.nl/
- https://huggingface.co/datasets/oscar
- https://grambank.clld.org/
- http://www.cs.cmu.edu/~dmortens/projects/7_project/
- https://wals.info/
- https://universaldependencies.org/
- https://aclanthology.org/L14-1215/

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- No

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or Data transfer agreements, Research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g. in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, codebook.tsv etc. where this information is recorded).

Developed software will be documented including README files.

Will a metadata standard be used to make it easier to find and reuse the data?
If so, please specify which metadata standard will be used.

If not, please specify which metadata will be created to make the data easier to find and reuse.

- No

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Other (specify below)

- Software: the implementation of our software will be released as open source on GitHub (https://github.com).
- Models trained will be made available on HuggingFace (https://huggingface.co/) including checkpoints of different epochs, so that researchers can analyse the training dynamics of our models.

How will the data be backed up?

- Other (specify below)

Everything is backed up on the flemish supercomputers as well as our local server which have daily backup procedures.

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

We have a dedicated storage space on the flemish supercomputer to which only us and the administrators have access.
We also have space on a local server to which again only our group has access.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

About 40EUR/year, this has been budgeted.

Data Preservation after the end of the Research Project

Which data will be retained for 10 years (or longer, in agreement with other retention policies that are applicable) after the end of the project?

In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

- All data will be preserved for 10 years according to KU Leuven RDM policy

Where will these data be archived (stored and curated for the long-term)?

- Other (specify below)

On a local server.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

No extra cost expected.

Data Sharing and Reuse

Will the data (or part of the data) be made available for reuse after/during the project?
Please explain per dataset or data type which data will be made available.

- Yes, as open data

Everything will be released on Github, as explained above.

If access is restricted, please specify who will be able to access the data and under what conditions.

NA

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Please explain per dataset or data type where appropriate.

- No

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other (specify below)

On our group's github.

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- MIT licence (code)

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, my dataset already has a PID

All our publications come with a DOI.

What are the expected costs for data sharing? How will these costs be covered?

No cost, we use the free version of github.

Responsibilities

Who will manage data documentation and metadata during the research project?

The first author of the papers.

Who will manage data storage and backup during the research project?

The first author of the papers.

Who will manage data preservation and sharing?

The first author of papers.

Who will update and implement this DMP?

The PI, with help from the researchers.