Datasheet for Dataset: Papyrus

Frédéric Piedboeuf
RALI, Diro
Université de Montréal
frederic.piedboeuf@umontreal.ca

Philippe Langlais
RALI, Diro
Université de Montréal
felipe@iro.umontreal.ca

1 Motivation

For what purpose was the dataset created? The dataset was created in order to address the problem of French, as well as multilingual, keyphrase generation.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The papyrus repository belongs to Université de Montréal (UdeM) which gave the RALI research lab of UdeM the right to prepare and distribute the present dataset for research purposes. The dataset has been prepared as part of the first author doctoral research.

Who funded the creation of the dataset? The dataset was created independently as part of the first author doctoral research, in cooperation with the Papyrus team from Université de Montréal.

Any other comments? This dataset was created and curated in partnership with the Papyrus team of Université de Montréal. We thank them for letting us use and publish this data, which will without a doubt be a great use to the research community.

2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Each instance represent an abstract, a title, and keyphrases associated with a document that was uploaded on Papyrus. Most of these documents are theses, but there are additional documents as well, such as institutional public reports, presentations, or documents uploaded by members of the faculty.

How many instances are there in total (of each type, if appropriate)? There are 26508 documents in the *non-curated* version of the dataset, among them 16249 valid ones. By *curated* we mean any preprocessing/filtering that was done after the collection step. Each document is identified by a unique index which corresponds to the index in Papyrus. Invalid documents include indexes that point to error pages, as well as documents that have no keyphrases, titles, or abstracts.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The dataset itself contains all documents from Papyrus, but we also release four curated subsets of it that do not contain all entries. Papyrus-f contains only French keyphrases and abstracts, and Papyrus-e only English ones. Papyrus-m entries are one abstract as well as associated keyphrases that have the same language, and Papyrus-a contains all data, with one entry being all abstracts and keyphrases of a document.

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

What data does each instance consist of? Each instance consist of a title, one or more abstracts in several languages, as well as several keyphrases as written by the author of the document.

Is there a label or target associated with each instance? Yes, the targets/labels are the keyphrases.

Is any information missing from individual instances? Some of the documents do not have keyphrases in all languages, which will lead to invalid entries when separated for the various tasks.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? In some subtasks, multiples translation of the same abstracts will find itself as multiple inputs. Other than that, some documents have been written by the same authors, for examples students who write more than one thesis. Also, professors typically supervise several students and therefore there is a possible relation between theses (and therefore abstracts) written by the students supervised by a same professor.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes, we provide the data already split.

Are there any errors, sources of noise, or redundancies in the dataset? Keyphrases are automatically assigned to a language using a simple heuristic: For keyphrases present in the abstract, we assign them to all the abstracts that contain them. For absent keyphrases, we use `fastText` and assign it to the most likely language. If `fastText` cannot identify the languages[1], such as for the keyphrase "間 ma", "1000", or "Leptoquark", we assign it to all abstracts for that entry. This causes some source of error, but we estimated an accuracy of 98.9% of accuracy on the keyphrases using 100 random entries from Papyrus-m. There is also possible sources of redundancies in the data. The same abstract and keyphrases were sometimes assigned to two or more documents. We remove the additional copies in the curated data.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is self-contained.

Does the dataset contain data that might be considered confiden- tial (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individ- uals' non-public communications)? No. All documents are published openly.

Does the dataset contain data that, if viewed directly, might be of- fensive, insulting, threatening, or might otherwise cause anxiety? Potentially : The dataset is composed of theses covering a wild range of subjects, including but not limited to rape, police brutality, or the n word.

Does the dataset identify any subpopulations (e.g., by age, gen- der)? Some instances of the dataset are theses discussing specific populations, such as women, queer people, or some other subpopulations.

Is it possible to identify individuals (i.e., one or more natural per- sons), either directly or indirectly (i.e., in combination with other data) from the dataset? Yes, one simply has to research the title or the abstract on a search engine to find the original thesis and the author(s) who wrote it.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual ori- entations, religious beliefs, political opinions or union member- ships, or locations; financial or health data;

[1] If the top-15 languages from `fastText` are not concordant with the languages of the abstracts.

biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No.

Any other comments? No.

3 Collection Process

How was the data associated with each instance acquired? Each document came in Papyrus with some meta-data that we used to extract the abstracts, titles, and keyphrases. Separation into languages for the various tasks was however conducted automatically with a process that we estimate accurate at 98.9%.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? A scrapping process was used to collect the data off the internet and then curated with heuristics as described before.

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

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? Only the authors of the papers were involved in the data collection process.

Over what timeframe was the data collected? The latest version of the dataset was collected on April 7, 2022. The whole scrapping and curating of the dataset took approximately two weeks.

Were any ethical review processes conducted (e.g., by an institu- tional review board)? No, but we asked and obtained permission from the people in charge of Papyrus for the collection and distribution of the dataset.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? We collected it from the Papyrus website.

Were the individuals in question notified about the data collec- tion? No, but the people in charge of Papyrus were notified and gave consent.

Did the individuals in question consent to the collection and use of their data? N/A.

If consent was obtained, were the consenting individuals pro- vided with a mechanism to revoke their consent in the future or for certain uses? N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been con- ducted? N/A.

Any other comments? No.

4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of miss- ing values)? Language assignment of both abstract and keyphrases was done using the method described earlier. Then when separating into the different tasks, some instances were dropped due to lack of labels. All original data is however provided with the dataset.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes, the raw data is available in a jsonlines format along the curated dataset.

Is the software that was used to preprocess/clean/label the data available? Yes, all preprocessing is available in the github[2].

Any other comments? No.

5 Uses

Has the dataset been used for any tasks already? The dataset has been used for keyphrases generation in a paper submitted to Neurips dataset 2022.

Is there a repository that links to any or all papers or systems that use the dataset? Not at the moment.

What (other) tasks could the dataset be used for? The dataset could be used in a reverse direction, that is to say for abstract or title generation from the keyphrases, as well as any downstream tasks that use keyphrases. One could also not use the keyphrases and do for example a task of title generation from the abstract or the reverse.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might im- pact future uses? No.

Are there tasks for which the dataset should not be used? No.

Any other comments? No.

6 Distribution

Will the dataset be distributed to third parties outside of the en- tity (e.g., company, institution, organization) on behalf of which the dataset was created? The dataset will be available to all for research purposes under the creative common license.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset will be hosted on the RALI website[3], and downloading instructions will be available on the github linked above.

When will the dataset be distributed? The dataset is available for download since may 2022.

Will the dataset be distributed under a copyright or other intel- lectual property (IP) license, and/or under applicable terms of use (ToU)? The dataset is distributed under the creative commons public license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

Any other comments? No.

7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The dataset will be hosted, supported, and maintained by the RALI team and the first author.

[2] https://github.com/smolPixel/French-keyphrase-generation

[3] http://rali.iro.umontreal.ca/rali/en

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The author can be contacted at frederic.piedboeuf@umontreal.ca

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If enough new instances are available that it becomes pertinent to create a new dataset and repeat the study, a follow-up will be made on this subject, and a new paper describing the differences between the old and new datasets submitted. If errors are detected, the dataset will be corrected and changes described on both the RALI website and the Github.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? No.

Will older versions of the dataset continue to be supported/hosted/maintained? Yes. In both cases or errors or follow up studies, the changed will be described on Github and the RALI website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? We actively encourage others to extend the dataset by getting more multilingual data. Please contact the first author if you have an extension to the dataset.

Any other comments? No.