Data sheet Documentation

These are the sections and questions you need to answer when submitting your datasheet. Please note, some questions might not apply to your dataset. If they do not you can delete the question or respond with 'No'. See this paper for more info.

Please provide as much information as possible about your data set.

Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

1. For what purpose was the dataset created?

   a. The main goal of this effort is to create a dataset usable for pretraining any kind of general purpose language model. A further benefit is to open up a reliably-updated topic-specific dataset in Kiswahili in two forms: the extracted text on zindi as a baseline for the language modeling tool behind tesseract-ocr's Swahili reading features, and as png images for further research in OCR.
2. Was there a specific task in mind?
   a. Pretraining language models and word vectors.
3. Was there a specific gap that needed to be filled?
   a. Lack of large contiguous unlabeled datasets of Swahili.
4. Who created this dataset
   a. Brian Muhia

Composition

1. What do the instances that comprise the dataset represent?
   a. Individual text files each containing the contents of a single page from the corpus of png images extracted by imagemagick. The language of the text is Kiswahili and was produced using tesseract 4.0, using this command:
      `tesseract target-png/$f extracted-txt/$f -l swa.` The "-l swa" flag indicates the usage of a Swahili language model used by tesseract to ease extraction of Swahili text.
2. Are there multiple types of instances
   a. No, just text
3. How many instances are there in total (of each type, if appropriate)? 2769 instances
4. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
   a. Dataset contains all possible instances at the moment.
5. What data does each instance consist of?
   a. "Raw" data (e.g., unprocessed text). We provide this in raw form in order to log

the outputs of the tools used before any attempts are made to improve them. This is intended to be a support structure for further research in this area

6. Does the dataset contain data that might be considered confidential?
    a. No
7. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No

Collection Process

1. How was the data associated with each instance acquired?
    a. The data was directly observable i.e. raw text.
2. Over what timeframe was the data collected?
    a. This was a crawl done in December 2019
3. What mechanisms or procedures were used to collect the data
    a. wget was used to download pages containing links to pdfs, and the pages were parsed to extract the links, which were then downloaded with a 2s latency to minimize cost on the hosting servers.
    b. How were these mechanisms or procedures validated? The rules were set based on research done on how best to extract the documents.
4. What was the resource cost of collecting the data?
    a. About 5 days of runtime to download, and 5 hours of compute to extract the documents from pdf-png and then png-txt.

5. Who was involved in the data collection process?
    a. Brian Muhia
6. Were any ethical review processes conducted?
    a. No. It does not involve individual people.

Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done?
    a. Some pdfs had errors in the imaging process, so all of the resulting png files were removed from the corpus.

2. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data?
    a. Yes, here https://drive.google.com/open?id=13Zn2x7OKogYcJMTghN-zCZnAFxHNoB05
3. Is the software used to preprocess/clean/label the instances available?
    a. https://github.com/tesseract-ocr/tesseract

Uses

1. Has the dataset been used for any tasks already? No
2. Is there a repository that links to any or all papers or systems that use the dataset?
    a. https://github.com/COCOHubCommunity/tz-hansard-extraction
3. What (other) tasks could the dataset be used for?
    a. A possible application could be autocorrection. An experiment with fasttext indicates that different spelling errors of a specific term are grouped together in some cases, but no clear measure of this tendency has been taken. Tests of usefulness have been done using fasttext, and it appears that in some cases words with transcription errors e.g. “tulipatannn” instead of “tulipatana” are grouped together which might be useful in spelling correction applications.

Distribution

1. Will the dataset be distributed to third parties outside of the entity?
    a. Anyone will have access to this dataset.
2. How will the dataset will be distributed?
    a. GitHub, Google Drive
3. Does the dataset have a digital object identifier (DOI)? 10.5281/zenodo.6636622
4. When will the dataset be distributed? Q2 2022
5. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
    a. Creative Commons Attribution v4.0
6. Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
    a. No
7. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
    a. No

Maintenance

1. Who is supporting/hosting/maintaining the dataset? Brian Muhia
2. How can the owner/curator/manager of the dataset be contacted?
    a. [redacted]
3. Is there an erratum? An issue list is available at github:
    a. Send an email with subject line “Errors RE: Swahili-Tz-Hansard Dataset”
4. Will the dataset be updated?
    a. Yes. The critical path for improvements to this dataset includes more research into the language models used by tesseract-ocr and how to improve them to get better OCR scans. Other than that, we may find possible improvements in

cropping strategies for the png images, and how to enable parallelization without bugs involving imagemagick.
    b. Except updates a few times a month (at least 3) on GitHub, at this repository: https://github.com/COCOHubCommunity/tz-hansard-extraction

5. Will older versions of the dataset continue to be supported/hosted/maintained?
    a. Yes, we will use cloud storage backups, and any older versions may be requested.
6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
    a. Contributing to the project through github: https://github.com/COCOHubCommunity/tz-hansard-extraction
    b. Will these contributions be validated/verified? Yes, through the research and data curation community masakhane.io
    c. Is there a process for communicating/distributing these contributions to other users? GitHub issues on the repository https://github.com/COCOHubCommunity/tz-hansard-extraction