Data sheet Documentation

Motivation

1. For what purpose was the dataset created?
   This dataset was created to provide Dhopadhola(ADH) to English Parallel sentences to help in availing services that require Natural Language Processing to Dhopadhola speakers. The dataset can also be used to study transfer learning in related African languages as it is closely related to Dholuo spoken in Kenya & Tanzania, Acholi, Lango and Alur in Uganda and other Luo languages.

2. Was there a specific task in mind?
   The dataset can be used for Machine Translation purposes. It consists of 2484 parallel (Dhopadhola and English) sentences from different domains and 3386 monolingual Dhopadhola sentences. Both Supervised and Semi-supervised MT can utilise this dataset.

3. Was there a specific gap that needed to be filled?
   Dhopadhola is a very low resourced language; it has very few resources available publicly on the internet and even in other print media. This dataset is will help in the availability of Dhopadhola in digital media as when the task for which it is intended for(Machine Translation) is implemented, more resources will be translated into the language and also the native speakers will be incentivized to use it online eg on social media because non-speakers can get the translations.

4. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization?
   I, Perez Ogayo , created this dataset by collecting sentences from different sources.

5. What support was needed to make this dataset? e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.
   I received the support of volunteer native speakers to check on the correctness and alignments of sentences and also translation of some sentences. The associated costs were incurred by me and the volunteers.

6. Any other comments?

Composition

1. What do the instances that comprise the dataset represent? e.g., documents, photos, people, countries?
   They comprise sentences in Dhopadhola and their translation in English collected across various sources and describing different things.

2. Are there multiple types of instances e.g., movies, users, and ratings; people and interactions between them; nodes and edges? Please provide a description.
   N/A

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

Source	Domain	Parallel/Monolingual	No of Sentences
Global Story Books	Fiction	Parallel	166
StoryWeaver stories	Fiction	Parallel	316
Corona Virus Poster	Medical	Parallel	17
Job Offer Letter	Legal	Parallel	9
Mathew 1-7 Bible	Religion	Parallel	372
Jehovah Witness website, books and brochures	Religion	Parallel	1582
Adhola similes and proverbs	Literature	Parallel	22
History and customs of Dhopadhola	History and Culture	Adhola	1464
Kisoma Dhopadhola	Language Education	Adhola	1124
Piascy	Health Education	Adhola	465
Adhola similes and proverbs	Literature	Adhola	27
Songs(Adhola Anthem)	History and Culture	Adhola	15
Social media	Mixed	Adhola	291
Total			**5870**

Table 1

4. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not

representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset consists of samples collected from various sources. For example, for the Bible, I only included Matthew 1-9.

5. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
   Each instance consists of text collected from the sources shown in *table 1* above.

6. Is there a label or target associated with each instance? If so, please provide a description.
   The first 2484 Adhola sentences have an English translation which is the target. The last 3386 of the instances consist of Monolingual Adhola sentences.

7. 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).

   Everything is included in the dataset

8. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
   None explicitly.

9. 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?
   No

10. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
    No

11. Does the dataset relate to people? If not, you may skip the remaining questions in this section.
    No

12. Any other comments?

Table 2 below describes some statistics of the dataset.

Total Size	5870
Parallel	2484
Parallel %	42.3

Monolingual(Adhola)	3386
Monolingual(Adhola) %	57.6

Table 2

Collection Process

1. How was the data associated with each instance acquired?
    a. Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

    Most of the sentences in this database were gathered from websites and books on the web using software to crawl websites. Some were also manually typed from images of documents that were shared on social media.

2. Over what timeframe was the data collected?
    a. Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

    The dataset was collected over a period of 6 weeks from June 15th to August 2nd. The timeframe doesn't match the creation time frame.
    The first data source was published in 1960 and the latest data source was published in June 2020.

3. What mechanisms or procedures were used to collect the data
    a. e.g., hardware apparatus or sensor, manual human curation, software program, software API

    I used Jupyter notebook, Python programming language with different python packages such as BeautifulSoup package to scrape the data from different websites and pandas to clean and organize the data.
    Manual human curation was also involved especially in getting data from books, images and translating English texts and aligning the sentences.

    b. How were these mechanisms or procedures validated?

    ` N/A
4. What was the resource cost of collecting the data?
    a. (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell et al. for approaches in this area.)

- High speed internet access
- A laptop with good computational power(Intel core i7, 8GB Ram)

5. If the dataset is a sample from a larger set, what was the sampling strategy?
   N/A

6. 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)?
   Me (Perez Ogayo), a rising 4th year Bachelor degree student. 2 Ugandan Bachelor's degree students.

7. Were any ethical review processes conducted?
   a. (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
      N/A

   b. Does the dataset relate to people? If not, you may skip the remainder of the questions in 7 this section.
      No

Preprocessing/cleaning/labeling

1. 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 missing value? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
   The following steps were taken to process the data:
   1. Removing duplicate sentences
   2. Fixing the symbol which is sometimes not rendered correctly depending on the encoding format. Where it was wrongly represented, I manually replaced it.
   3. Removing non-Adhola(Mostly English words) which appear in brackets in the sentences.
   4. For the parallel sentences, manually making sure that they were aligned, if not align them.
2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?e.g., to support unanticipated future uses? If so, please provide a link or other access point to the "raw" data.
   Yes. They will be accessed from the project's GitHub repository.

3. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
   Pandas and Microsoft Excel

4. Any other comments?

Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.
   No

2. Is there a repository that links to any or all papers or systems that use the dataset?If so, please provide a link or other access point.
   No

3. What (other) tasks could the dataset be used for?
   When the Dhopadhola sentences are labeled appropriately, they can be used for text classification, text summarization and topic modeling

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
   a. For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description.
      The data set contains data from the Jehovah Witness website and the Bible and may contain Christian and Jehovah Witness views and opinions.

   b. Is there anything a future user could do to mitigate these undesirable harms?
      Unknown

5. Are there tasks for which the dataset should not be used?If so, please provide a description.
   The dataset should not be used for tasks that are high stakes (e.g. law enforcement)

6. Any other comments?

Distribution

1. Will the dataset be distributed to third parties outside of the entity? e.g., company, institution, organization on behalf of which the dataset was created? If so, please provide a description.
   Yes. The dataset will be made publicly available

2. How will the dataset be distributed?
   The dataset will be distributed via Github Repository.

3. Does the dataset have a digital object identifier (DOI)?
   No

4. When will the dataset be distributed?
   The dataset will be available in the Github Repository around September 2020.

5. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
   The dataset will be distributed under Creative Commons CC BY 4.0 License.

6. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
   There are no fees or restrictions

7. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
   Unknown

Maintenance

1. Who is supporting/hosting/maintaining the dataset?
   I will be maintaining the dataset (Perez Ogayo)

2. How can the owner/curator/manager of the dataset be contacted?e.g., email address
   All questions and comments can be sent to Perez Ogayo: perezogayo@gmail.com

3. Is there an erratum? If so, please provide a link or other access point.
   All changes to the dataset will be listed in the dataset's Github repository.

   Will the dataset be updated?e.g., to correct labeling errors, add new instances, delete instances.If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset will be updated to correct labeling spelling errors, misalignments and add new sentences. I will be responsible for updating the dataset. All changes to the dataset will be listed in the dataset's Github repository.

4. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted? If so, please describe these limits and explain how they will be enforced.
   N/A

5. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
   The older version of the dataset will continue to be supported, hosted and maintained via its GitHub repository.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description.
   Yes, I welcome others to extend and contribute to the dataset. If anyone is interested they can make a pull request with their updates on the dataset's GitHub repository.

   Will these contributions be validated/verified? If so, please describe how. If not, why not?
   They will be validate and verified by experienced translators and native speakers through GitHub and the project's Doccano site.

   Is there a process for communicating/distributing these contributions to other users?
   They will be listed on the project's GitHub repository

Implementation Plan

1. An explanation on how you would expand this data set if you won the $1,500 research grant.

   I seek to expand the dataset to 100,000 parallel sentences. I will work with native speakers, language teachers and translators to ensure accuracy of the sentences and a well-balanced representation of different domains and data sources. I am already working with a community of Dhopadhola speakers online who have offered to help me expand the dataset. I have also set up a Doccano site to facilitate the expansion process.

2. What steps would you take? What would be an approximate work plan?

Steps

1. Have a training and alignment meeting with the volunteers to discuss the steps and process
2. Identify and group the validators and translators o
3. With the help of the team, identify the news sources and news, legal documents and medical documents that will be translated to Dhopadhola
4. Assign the different domains to the validators
5. Start the collection, translation and validation process

Workplan

Milestone	Duration
• Verify and validate existing sentences and expand to 10,000 parallel sentences. • Training of volunteers	1 months
Collect, translate and validate parallel sentences from: • news sources - 20,000 • medical domain -20,000 • legal domain - 20,000	3.5 months
Collect, translate and validate 20,000 sentences from the FloRes Wikipedia English sentences	1.5 months
• Collecting 10,000 parallel sentences from social media and other informal chat platforms, • Dataset administaration work- set up newsletter , website etc	1 month
Total	7 months

3. What would the developed dataset finally look like?

The dataset will have 100,000 Adhola-English Parallel sentences. The sentences will be balanced across major domains and sources. The sentences will also have been validated by professional Dhoadhola translators.