TeDDi: Crawling issues

Olga Pelloni (Sozinova)

October 1, 2022

In this document, I describe the known issues in the crawled TeDDi files. In addition to the issues, I suggest potential ways of fixing them and report on the bugs I already fixed.

1 Mixed languages

It is not always purely one language per text file. In case of OpenSubtitles, we notice sometimes double subtitles in English in parallel to the main language (often in case of Mandarin).

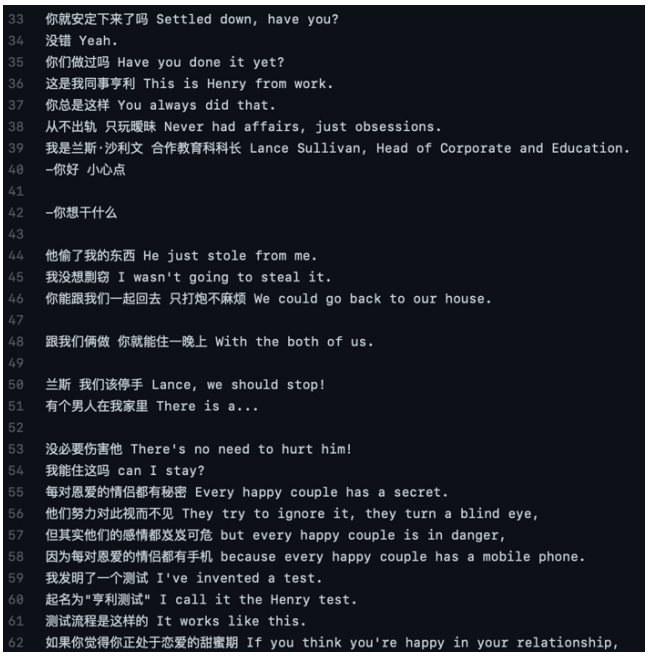

Figure 1: Several languages in one text file.

In case of Gutenberg texts, we notice some texts to be dictionaries, thus, containing text data in two languages as well. We try to avoid them while crawling.

Potential solution:

- Potential language identification tool `https://polyglot.readthedocs.io/en/latest/`

1.1 Solving the issue

I've tried using the Polyglot identification tool. It returns the relative number of up to 3 languages identified in a text. I set the following threshold: language 1 less than 99%, language 2 more than 5%. The results are written down in a text file.

Most of the texts seem to mix together with English, but not always. There are less obvious cases, such as Chinese - Spanish, Finnish - Swedish, Chinese simplified - Chinese traditional. Next step is to look closer at this results file and to decide what to do with the text files: keep the most present language inside a file or remove the file. Note that mixed languages can appear also in a single line, which means that if we want to clean the file, we have to identify a language token by token, which might not work that well.

Some files in the results belong to low-resource languages, so they probably have languages identified by mistake. It is better to filter out my text file only to the crawled languages.

2 Non-linguistic noise

Any non-linguistic items, such as tags (HTML, XML), some Latex-like commands, etc.

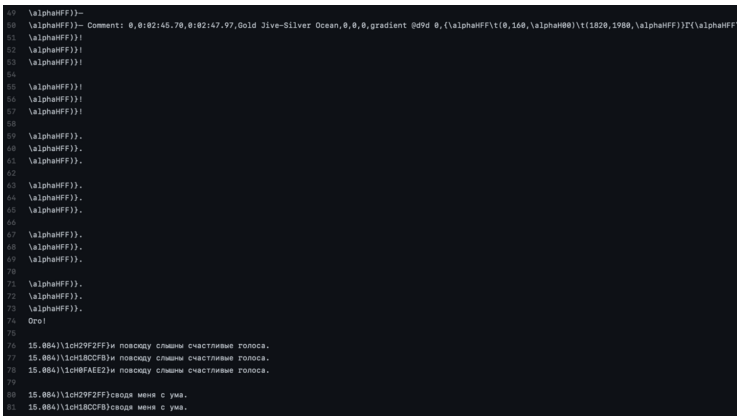

Figure 2: Non-linguistic noise example.

Potential solution:

- Identify the least frequent symbols or N-grams, look at the context of those symbols

2.1 Solving the issue

I've implemented an algorithm, which creates a frequency dictionary of the characters in a text file, filters out all the characters which have a frequency lower than 20.

Then, we search for sequences in a text of length 5, where every character has a frequency lower than 20. These sequences are potential non-linguistic noise strings.

This is a results file, where I write down a filename and an identified sequence. Many of them point out to mixed languages (mostly parts of English words), rather than to noise. A potential way to filter this file, is to look at the sequences containing non-letter symbols (punctuation marks, slashes, etc.). However, some sequences containing letters, and yet belong to noise, e.g.: `heb_nfi_1064.txt, 080ff`, which is a color code, included in an HTML tag. Quotation marks might help with such cases.

Once we find a way to filter out the actual noise in the results file, we can dig into that file and look at the context, in order to clean the noise. This, most probably, will require manual cleaning (the context can be the whole line, part of the line, whole tag or part of a tag, etc), or, maybe, the most frequent noise items can be cleaned out automatically (e.g. `color = "0FFFFF"` and alike).

3 Malformed sentences

Strange appearance of the letters, Unicode problems (symbols are not combined together as they should). Example of a Thai text:

Figure 3: Malformed sentences in Thai.

Potential solution:

- Check the Unicode sequence somehow (?), Unicode decomposition

3.1 Solving the issue

After looking closer at the suspicious Thai files, it turns out that Unicode is fine there, however, the text itself contains nonsense words and many control characters (non-printing characters), which is probably a sign of a strange encoding or machine, where the text was created. We found such files by filtering the segmentations obtained with BPE-200, length of words > 30 (with very regular segments' lengths).

The list of corrupted files: tha_nfi_67.txt, tha_nfi_688.txt, tha_nfi_915.txt. These files are to be deleted.

The file tha_nfi_46.txt has to be filtered from noise (tag "/c.bg_transparent"), otherwise, the text is fine.

Status: DONE (pull request accepted)

4 Meta in the text body

In some text files from the Gutenberg Project, there is meta information inside the text body. It often includes either the start or the ending of the Gutenberg copyright information, e.g. `***START OF THE PROJECT GUTENBERG EBOOK IT WAS MARLOWE***`.

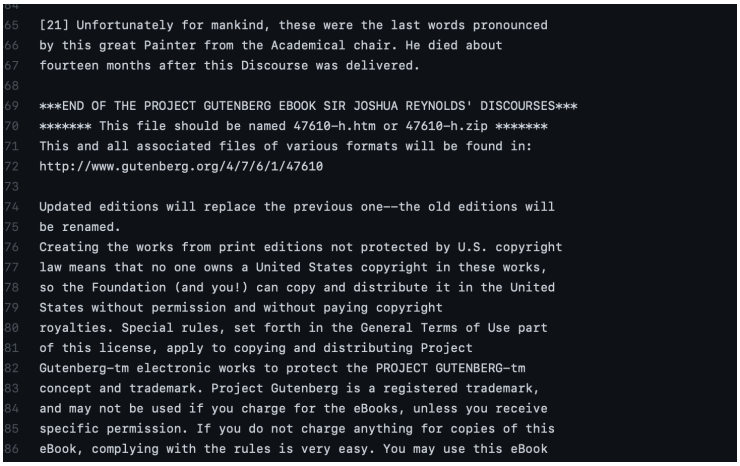

Figure 4: Meta related information in the text body.

Potential solution:

- Find the key phrases (Start/End of the Project Gutenberg...), remove the text before `START` and after `END`.

Files: eng_fic_97.txt, eng_fic_54.txt, eng_fic_68.txt, eng_fic_57.txt, eng_fic_92.txt, eng_fic_51.txt, eng_fic_50.txt, eng_fic_23.txt, eng_fic_20.txt, eng_fic_32.txt, eng_fic_17.txt, eng_fic_15.txt, eng_fic_8.txt, eng_fic_12.txt, eng_fic_48.txt, eng_fic_89.txt, eng_fic_67.txt, spa_fic_63.txt, spa_fic_66.txt, spa_fic_67.txt, spa_fic_108.txt, spa_fic_104.txt, spa_fic_23.txt, spa_fic_111.txt, spa_fic_35.txt, spa_fic_31.txt, spa_fic_96.txt, fra_fic_84.txt, fra_fic_79.txt, fra_fic_78.txt, fra_fic_83.txt, fin_fic_9.txt, fin_fic_120.txt, fin_fic_136.txt, fin_fic_10.txt, fin_fic_38.txt, fin_fic_123.txt, fin_fic_132.txt, fin_fic_118.txt, fin_fic_142.txt, fin_fic_71.txt, fin_fic_73.txt, fin_fic_140.txt, fin_fic_62.txt, fin_fic_151.txt, fin_fic_147.txt, fin_fic_153.txt, fin_fic_61.txt, fin_fic_79.txt, fin_fic_160.txt, fin_fic_47.txt, fin_fic_91.txt, fin_fic_159.txt, fin_fic_80.txt, fin_fic_82.txt, fin_fic_83.txt, fin_fic_27.txt, fin_fic_115.txt, fin_fic_101.txt, fin_fic_129.txt, fin_fic_19.txt, fin_fic_21.txt, fin_fic_7.txt, fin_fic_6.txt, fin_fic_107.txt, fin_fic_4.txt, fin_fic_23.txt, fin_fic_138.txt, fin_fic_110.txt, deu_fic_134.txt, deu_fic_36.txt, deu_fic_18.txt, deu_fic_118.txt, deu_fic_54.txt, deu_fic_96.txt, deu_fic_43.txt, deu_fic_151.txt, deu_fic_52.txt, deu_fic_44.txt, deu_fic_74.txt, deu_fic_17.txt, deu_fic_111.txt

4.1 Solving the issue

Removing meta info file by file. There is sometimes a long text of a full Gutenberg license. It might be a good idea to rerun the reports after such cleanings (line count, tokens).

In some files I modify the meta field short_description, if the author and title were missing; the same applies to the field year_published.

General observation: the issue often appeared in the samples, where we sample text in the end of a novel, and then continue from the beginning. The crawler could be improved by taking into account this scenario, and cleaning the license in the end and the meta in the beginning. However, this was also the case in the "whole" type of samples.

Note: some files have a tab in the short_description field. Maybe it needs to be checked.

Files done:

eng_fic_97.txt, eng_fic_54.txt, eng_fic_68.txt, eng_fic_57.txt, eng_fic_92.txt, eng_fic_51.txt, eng_fic_50.txt, eng_fic_23.txt, eng_fic_20.txt, eng_fic_32.txt, eng_fic_17.txt, eng_fic_15.txt, eng_fic_8.txt, eng_fic_12.txt, eng_fic_48.txt, eng_fic_89.txt, eng_fic_67.txt

Pull request done, will be continued with the other files.

UPD 30.09.22:

Finished all the files (pull request finished too):

Spanish:

spa_fic_63.txt, spa_fic_66.txt, spa_fic_67.txt, spa_fic_108.txt, spa_fic_104.txt, spa_fic_23.txt, spa_fic_111.txt, spa_fic_35.txt, spa_fic_31.txt, spa_fic_96.txt

Finnish:

fin_fic_9.txt, fin_fic_120.txt, fin_fic_136.txt, fin_fic_10.txt, fin_fic_38.txt, fin_fic_123.txt, fin_fic_132.txt, fin_fic_118.txt, fin_fic_142.txt, fin_fic_71.txt, fin_fic_73.txt, fin_fic_140.txt, fin_fic_62.txt, fin_fic_151.txt, fin_fic_147.txt, fin_fic_153.txt, fin_fic_61.txt, fin_fic_79.txt, fin_fic_160.txt, fin_fic_47.txt, fin_fic_91.txt, fin_fic_159.txt, fin_fic_80.txt, fin_fic_82.txt, fin_fic_83.txt, fin_fic_27.txt, fin_fic_115.txt, fin_fic_101.txt, fin_fic_129.txt, fin_fic_19.txt, fin_fic_21.txt, fin_fic_7.txt, fin_fic_6.txt, fin_fic_107.txt, fin_fic_4.txt, fin_fic_23.txt, fin_fic_138.txt, fin_fic_110.txt

German:

deu_fic_134.txt, deu_fic_36.txt, deu_fic_18.txt, deu_fic_118.txt, deu_fic_54.txt, deu_fic_96.txt, deu_fic_43.txt, deu_fic_151.txt, deu_fic_52.txt, deu_fic_44.txt, deu_fic_74.txt, deu_fic_17.txt, deu_fic_111.txt

Comments to the files:

- fra_fic_84.txt, fra_fic_79.txt, fra_fic_78.txt, fra_fic_83.txt: look fine, no changes done; gutenberg is used as a word;
- spa_fic_108.txt, spa_fic_31.txt: a lot of English comments and there is a Spanish-English dictionary at the end of texts;
- deu_fic_44.txt: looks like Old German text or as dialectal German variety;

- Maybe it's better to explore NBSP symbols (like whitespace). I removed them in Spanish files (substituted with whitespace), in the other language files there are many NBSP symbols, I didn't change them. I think they need a check, how they would influence preprocessing in Python.

Status: DONE (pull request waiting for approval)

5 Missing meta information

In some text files from the Gutenberg Project, there is no value in the `short_description` meta field, which usually includes the author name and the title.

```
# language_name_wals:   English
# language_name_glotto: English
# iso639_3:     eng
# year_composed:        NA
# year_published:       2006
# mode: written
# genre_broad:  fiction
# genre_narrow: general_fiction
# writing_system:       Latn
# special_characters:   NA
# short_description:    NA
# source:       http://www.gutenberg.org/files/8600/8600-h/8600-h.htm
# copyright_short:      http://www.gutenberg.org
# copyright_long:       http://www.gutenberg.org
# sample_type:  part
# comments:     NA

Gervaise flushed. She thought she would have felt less shame if he had
 taken her in his arms and kissed her. Goujet was an odd fellow, proposing
 to elope, just the way it happens in novels. Well, she had seen plenty of
 workingmen making up to married women, but they never took them even as
 far as Saint-Denis.
```

Figure 5: Missing meta information (author name and title).

Potential solution:

- Search for the author name and the title at the HTML page

Files:

- eng_fic_108.txt
- eng_fic_106.txt
- eng_fic_107.txt
- spa_fic_17.txt
- spa_fic_22.txt
- spa_fic_21.txt

- spa_fic_20.txt
- spa_fic_18.txt
- spa_fic_19.txt

5.1 Solving the issue

Found the name of author and the title on the web pages. These samples were taken from two novels: Émile Zola, L'Assommoir and Leopoldo Alas, La Regenta. Copy-pasted this info manually.

Status: DONE (pull request accepted)

6 Manually added data

Consistency of tags (what goes to which layer, glosses or not in the segmentation layer), usage of symbols (asterix as a sound in Piraha), other consistency issues.

7 Status on 30.09.2022

Here is an update about the TeDDi cleaning status as of 30.09.2022.

- **Mixed languages**: I did not change anything in the TeDDi github repo. What should be done here next: The file should be filtered; 1) leave there only the crawled languages (remove low-resouce languages that we collected manually); 2) find empirically a threshold for confidence and filter according to that. For example, the file kor_nfi_271.txt has Korean with confidence 52 and English with confidence 47. Most likely, this is a good candidate to remove from the corpus. On the other hand, kor_nfi_275.txt has Korean with confidence 94 and English with confidence 5; probably, such file can be left unchanged.
- **Non-linguistic noise**: I did not change anything in the TeDDi github repo. What should be done here next: The results file has to be manually processed and filtered as well. This task will take more time than the mixed languages problem. First, one needs to leave only crawled languages as in the mixed languages problem. Then, one needs to go through all examples and identify which ones belong to noise and which do not. Some examples of how this could be automatized are given above in the respective section.
- **Malformed sentences**: Done in Thai, pull request accepted.
- **Meta in the text body**: Done in 4 languages, pull request waiting for approval. See comments left above in the respective section for potential things to check later.
- **Missing meta information**: Done in English and Spanish, pull request accepted.

- **Manually added data**: To be done later, probably after consulting Chris and Steve.