Reproducibility review of: Enhancing toponym identification: Leveraging Topo-BERT and open-source data to differentiate between toponyms and extract spatial relationships

Rémy Decoupes

2024-05-27

This report is part of the reproducibility review at the AGILE conference. For more information see https://reproducible-agile.github.io/. This document is published on OSF at https://osf.io/nbk57/. To cite the report use

> Decoupes R, (2024, April). Reproducibility review of: Enhancing toponym identification: Leveraging Topo-BERT and open-source data to differentiate between toponyms and extract spatial relationships. https://doi.org/10.17605/OSF.IO/nbk57

Reviewed paper

Shingleton J., and Basiri, A.: Enhancing toponym identification: Leveraging Topo-BERT and open-source data to differentiate between toponyms and extract spatial relationships, AGILE GIScience Ser., 5, 12, https://doi.org/10.5194/agile-giss-5-12-2024, 2024

Summary

As indicated in the Data and Software Availability section, the authors shared their code, data, and trained models through an OSF (Open Science Framework) repository. Through 4 notebooks, we were able to train two baseline models and then create a new training dataset to train the proposed model by the authors. These models were then compared with human evaluation (through shared data).

Evaluating the reproducibility of this article was not an easy task. In fact, this processing chain requires a lot of computational resources and time for its execution. Another difficulty was that the notebooks and Python library developed by the authors and shared via OSF contained some errors.

However, the authors accompanied me throughout this process, providing new versions of the code files to correct the errors I encountered. My feeling is that the reproducibility review process was beneficial. The scientific article was almost entirely reproduced.

Reproducibility reviewer notes

1. Installation

The authors shared their code using OSF as described in the Data and Software Availability Section. To retrieve it, I downloaded a zip file of their shared directory and I unzipped it. As mentioned in the README.md file the use of a GPU is recommended, so the reproducibility review was done on a server with Nvidia V100 (Ubuntu 20.04).

To set up the development environment, I used conda and installed the recommended versions of packages as listed in the README.md file.

```
conda create -n "agile-19" python=3.10.13 ipython pip ipykernel
conda activate agile-19

python -m ipykernel install --user --name=agile-19

pip install numpy==1.26.0
pip install pandas==2.0.3
pip install scikit-learn==1.3.0
pip install transformers==4.32.1
pip install tokenizers==0.13.0
pip install tensorflow==2.10.0
pip install shapely==2.0.1
pip install geopy==2.4.0
pip install wikipedia==1.4.0
pip install matplotlib
pip install tqdm
```

However, I noticed that the GPU from the server was not used. The authors assisted me in re-installing and configuring TensorFlow, but I was unsuccessful Here are the details of what I tried:

1. Installing on my own server

```
conda install -c conda-forge cudatoolkit=11.8.*
pip install nvidia-cudnn-cu11
pip uninstall tensorflow
pip install tensorflow[and-cuda]
```

But when I tried to list the devices that can be used (`print(len(tf.config.list_physical_devices('GPU')))`), only the CPU was shown.

2. Using tensorflow docker images

```
sudo docker run -it -p 8888:8888 -v ./:/tf tensorflow/tensorflow:latest-gpu-jupyter
```

Inside this container, I could run the code provided by the authors, but tensorflow still could not identify my GPU. So I decided to stop trying to make tensorflow work with cuda and instead ran the training on the CPU

2. Run the notebooks

The authors provided 4 notebooks which must be run sequentially (as explained in the README.md):

- NB1_training_TopoBERT_on_NER_data.ipynb
- NB2_extracting_wikipedia_data.ipynb
- NB3_relational_retagging.ipynb
- NB4_training_on_relationally_tagged_data.ipynb

2.1 NB1_training_TopoBERT_on_NER_data

When I ran this notebook, I encountered the following errors. The authors helped me by providing new files through their OSF repository. Here are the details:

- In section 1.1: there are typos on the data paths for train and test. Example with train: train_coNLL_dataset.csv instead of CoNLL_train_dataset.csv. I had to changed the order of the import of BERT_geoparser modules: `from train_model import Trainer` has to be the last import.

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
Cell In[1], line 8
      6 from sklearn.utils import class_weight
      7 # local imports
```

```
----> 8 from train_model import Trainer
      9 from tokenizer import Tokenizer
     10 from data import Data

File ~/agile/19/notebooks/../BERT_geoparser/train_model.py:13
     11 import tensorflow as tf
     12 # local imports
---> 13 from BERT_geoparser.tokenizer import Tokenizer
     14 from BERT_geoparser.data import Data
     15 from BERT_geoparser.model import BertModel

ModuleNotFoundError: No module named 'BERT_geoparser'
```

Author provide new files for their module. They fix this issue

- Error In first cell: `AttributeError: partially initialized module 'charset_normalizer' has no attribute 'md__mypyc' (most likely due to a circular import).` I had to force the reinstallation of the charset library.: `pip install --force-reinstall charset-normalizer==3.1.0`.

– Error on Cell 2, l.8 : `FileNotFoundError: [Errno 2] No such file or directory: 'r../models/TopoBERT_CoNLL`

```
trainer.train(save_as=’r../models/TopoBERT_CoNLL.hdf5’,
                n_epochs=20,
                batch_size=4,
                val_split=0.1)
```

There is a typographical mistake `'r../models/TopoBERT_CoNLL.hdf5'` instead of r‘../models/TopoBERT__CoNLL.hdf5’‘

- Because I was only running the notebook on CPU, I was not able to train the 2nd model (on WikiNeural). I asked the authors to upload their model `TopoBERT_WikiNeural.hdf5`. The authors told me they would upload this model but at the time of writing it was not yet available on the OSF repository.

2.2 NB2__extracting__wikipedia__data

Here are the execution errors I encountered:

- in cell 28, an import of library `nltk` was not in the list of requirements.

```
pip install nltk
```

- In cell 7 after section `2.3 Adding US Counties`, l.3: `location = remove_parentheses(row.County + ' County')`: raises error `AttributeError: 'Series' object has no attribute 'County'`. There is a typo, in the dataframe poly__counties, county column is in lowercase. I changed to `location = remove_parentheses(row.county + ' County')`.

- In the same cell, there is a similar issue with `State` instead of `state_name` in l.4

- Section 2.4: I was not able to run this section because I did not train the TopoBERT__WikiNeural.hdf5. The authors have told me that they will upload this model so it should be easy for a re-user to run these last cells.

2.3 NB3__relational__retagging

- The first run lead to an TypeError (`Retagger.retag() got an unexpected keyword argument 'offset'`) from the 2nd cell (in line l.5). The authors provide an update version of this notebook. This updated version need to install a new librairy `pyproj`.

```
pip install pyproj
```

- In cell 2, For every location (i.e. 6487), a warning is raised, like for example `Unable to search query "madrid"`

- In cell 3, same warnings for the 23299 locations. Authors updated the OSF repository and uploaded a correct wiki__cities__tagged.csv and an updated BERT__geoparser module.

- In cell 5, there is a typo in the name of a file (l.2: `wiki_counties = pd.read_csv('../data/NB3/**wiiki**_count`

- In cell 5, l.15, `wiki_counties = pd.read_csv('../data/NB3/wiiki_counties_retagged.csv')` raises an error: `TypeError: Population must be a sequence. For dicts or sets, use sorted(d)`

Authors update `data/NB2/wiki_cities_tagged.csv`

- In cell 8, after the update of `data/NB2/wiki_cities_tagged.csv`, I had a warning for each location `Unable to search query slovakia`

Authors update `BERT-geoparser/BERT-geoparser.py`

- In cell 8, after the update of `BERT-geoparser/BERT-geoparser.py`, I had another warning

```
Warning: invalid value encountered in intersection
  return lib.intersection(a, b, **kwargs)
```

After checking intermediate results with the Authors, we decided that I ignore these warnings.

- In cell 8, l.2, `reviewed_data = pd.read_csv('../data/NB3/wiki_places_reviewed_new.csv')` raises an error: `FileNotFoundError: [Errno 2] No such file or directory: '../data/NB3/wiki_places_re` Indeed, the file `wiki_places_reviewed_new.csv` is missing. I commented this line since it overwrites the first line.
- In cell8, l.6, `reviewed_data = replace_non(reviewed_data)` raises an error: `KeyError: 'Tag'`. Indeed, there is no `Tag` column in reviewed_data dataframe.

Authors update the 3rd notebook. With this new file, I was able to run all the notebook without error and produce the @fig1_1.

```
knitr::include_graphics("./figs/fig_1.png")
```

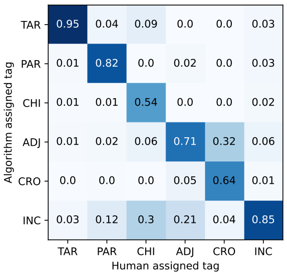


Figure 1: Reproduction of Figure 2

2.4 NB4_training_on_relationally_tagged_data

- Since I ran out of time, I decided to not retrain the proposed model `TopoBERT_relational.hdf5` as it was shared by the authors. I skipped the 2nd cell
- 3rd and 4th cell after section 4.1 Analyzing the results against test data: They print the 2 first columns of Table 3 `Nalg and Nmod`. But it seems to me that there are a few report inversions.

```
knitr::include_graphics("./figs/error_tab3.png")
```

- Cell 8: provide 2 confusion matrices

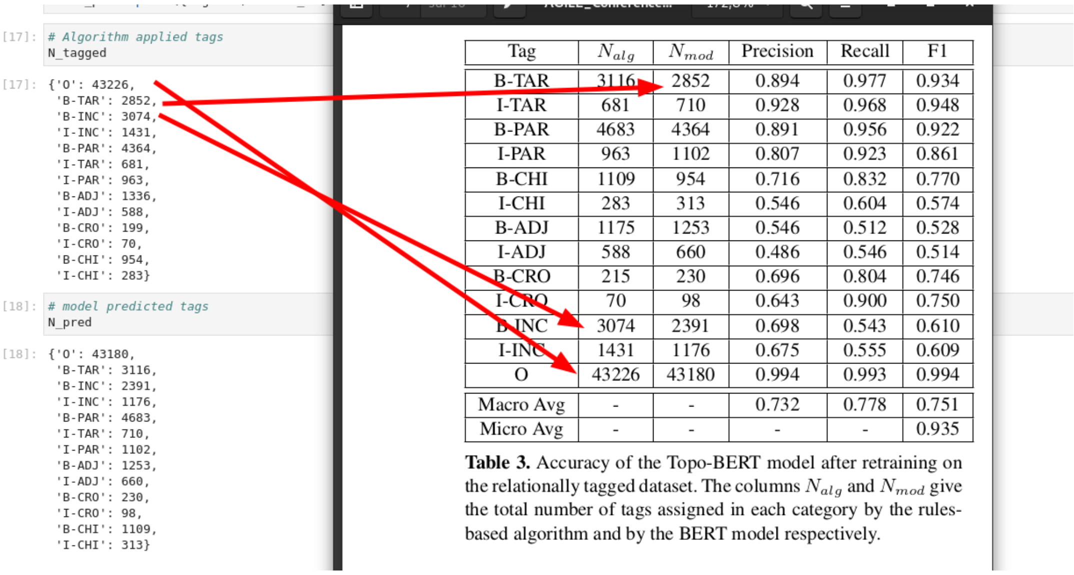
```
knitr::include_graphics("./figs/fig_2.png")
```

- In section 4.3 Precision, Recall and F1:

```
AttributeError: 'DataFrame' object has no attribute 'Tag'
```

- 2nd cell of section 4.3, complete **Table 3** with Precision, Recall, F1

```
B-TAR
Total assigned by human : 2852
Total assigned by algorithm : 3116
Precision : 0.894
Recall : 0.977
F1 : 0.934
###################################
```



```
[17]: # Algorithm applied tags
      N_tagged
```

```
[17]: {'O': 43226,
       'B-TAR': 2852,
       'B-INC': 3074,
       'I-INC': 1431,
       'B-PAR': 4364,
       'I-TAR': 681,
       'I-PAR': 963,
       'B-ADJ': 1336,
       'I-ADJ': 588,
       'B-CRO': 199,
       'I-CRO': 70,
       'B-CHI': 954,
       'I-CHI': 283}
```

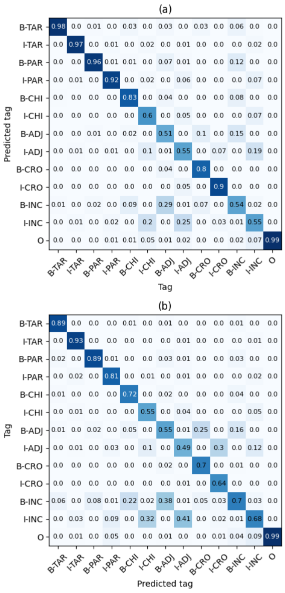
```
[18]: # model predicted tags
      N_pred
```

```
[18]: {'O': 43180,
       'B-TAR': 3116,
       'B-INC': 2391,
       'I-INC': 1176,
       'B-PAR': 4683,
       'I-TAR': 710,
       'I-PAR': 1102,
       'B-ADJ': 1253,
       'I-ADJ': 660,
       'B-CRO': 230,
       'I-CRO': 98,
       'B-CHI': 1109,
       'I-CHI': 313}
```

Tag	N_{alg}	N_{mod}	Precision	Recall	F1
B-TAR	3116	2852	0.894	0.977	0.934
I-TAR	681	710	0.928	0.968	0.948
B-PAR	4683	4364	0.891	0.956	0.922
I-PAR	963	1102	0.807	0.923	0.861
B-CHI	1109	954	0.716	0.832	0.770
I-CHI	283	313	0.546	0.604	0.574
B-ADJ	1175	1253	0.546	0.512	0.528
I-ADJ	588	660	0.486	0.546	0.514
B-CRO	215	230	0.696	0.804	0.746
I-CRO	70	98	0.643	0.900	0.750
B-INC	3074	2391	0.698	0.543	0.610
I-INC	1431	1176	0.675	0.555	0.609
O	43226	43180	0.994	0.993	0.994
Macro Avg	-	-	0.732	0.778	0.751
Micro Avg	-	-	-	-	0.935

Table 3. Accuracy of the Topo-BERT model after retraining on the relationally tagged dataset. The columns N_{alg} and N_{mod} give the total number of tags assigned in each category by the rules-based algorithm and by the BERT model respectively.

Figure 2: Example of inversion between the notebook and the manuscript: B-TAR from N-tagged in the notebook should be in the column $N(alg)$ of the manuscript.



(a)

Predicted tag	B-TAR	I-TAR	B-PAR	I-PAR	B-CHI	I-CHI	B-ADJ	I-ADJ	B-CRO	I-CRO	B-INC	I-INC	O
B-TAR	0.98	0.0	0.01	0.0	0.03	0.0	0.03	0.0	0.03	0.0	0.06	0.0	0.0
I-TAR	0.0	0.97	0.0	0.01	0.0	0.02	0.0	0.01	0.0	0.0	0.0	0.02	0.0
B-PAR	0.0	0.0	0.96	0.01	0.01	0.0	0.07	0.01	0.0	0.0	0.12	0.0	0.0
I-PAR	0.0	0.01	0.0	0.92	0.0	0.02	0.0	0.06	0.0	0.0	0.0	0.07	0.0
B-CHI	0.0	0.0	0.0	0.0	0.83	0.0	0.04	0.0	0.0	0.0	0.08	0.0	0.0
I-CHI	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.05	0.0	0.0	0.0	0.07	0.0
B-ADJ	0.0	0.0	0.01	0.0	0.02	0.0	0.51	0.0	0.1	0.0	0.15	0.0	0.0
I-ADJ	0.0	0.01	0.0	0.01	0.0	0.1	0.01	0.55	0.0	0.07	0.0	0.19	0.0
B-CRO	0.0	0.0	0.0	0.0	0.0	0.0	0.04	0.0	0.8	0.0	0.0	0.0	0.0
I-CRO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.05	0.0	0.9	0.0	0.0	0.0
B-INC	0.01	0.0	0.02	0.0	0.09	0.0	0.29	0.01	0.07	0.0	0.54	0.02	0.0
I-INC	0.0	0.01	0.0	0.02	0.0	0.2	0.0	0.25	0.0	0.03	0.01	0.55	0.0
O	0.0	0.0	0.0	0.01	0.01	0.05	0.01	0.02	0.0	0.0	0.02	0.07	0.99

(b)

Tag	B-TAR	I-TAR	B-PAR	I-PAR	B-CHI	I-CHI	B-ADJ	I-ADJ	B-CRO	I-CRO	B-INC	I-INC	O
B-TAR	0.89	0.0	0.0	0.0	0.01	0.0	0.01	0.0	0.0	0.0	0.01	0.0	0.0
I-TAR	0.0	0.93	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.01	0.0
B-PAR	0.02	0.0	0.89	0.01	0.0	0.0	0.03	0.01	0.0	0.0	0.03	0.0	0.0
I-PAR	0.0	0.02	0.0	0.81	0.0	0.01	0.0	0.01	0.0	0.0	0.0	0.02	0.0
B-CHI	0.01	0.0	0.0	0.0	0.72	0.0	0.02	0.0	0.0	0.0	0.04	0.0	0.0
I-CHI	0.0	0.01	0.0	0.0	0.0	0.55	0.0	0.04	0.0	0.0	0.0	0.05	0.0
B-ADJ	0.01	0.0	0.02	0.0	0.05	0.0	0.55	0.01	0.25	0.0	0.16	0.0	0.0
I-ADJ	0.0	0.01	0.0	0.03	0.0	0.1	0.0	0.49	0.0	0.3	0.0	0.12	0.0
B-CRO	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.0	0.7	0.0	0.01	0.0	0.0
I-CRO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.64	0.0	0.0	0.0
B-INC	0.06	0.0	0.08	0.01	0.22	0.02	0.38	0.01	0.05	0.03	0.7	0.03	0.0
I-INC	0.0	0.03	0.0	0.09	0.0	0.32	0.0	0.41	0.0	0.02	0.01	0.68	0.0
O	0.0	0.01	0.0	0.05	0.0	0.0	0.01	0.0	0.0	0.01	0.04	0.09	0.99

Figure 3: Reproduction of Figure 3 (a) and (b)

```
I-TAR
Total assigned by human : 681
Total assigned by algorithm : 710
Precision : 0.928
Recall : 0.968
F1 : 0.948
###################################
B-PAR
Total assigned by human : 4364
Total assigned by algorithm : 4683
Precision : 0.891
Recall : 0.956
F1 : 0.922
###################################
I-PAR
Total assigned by human : 963
Total assigned by algorithm : 1102
Precision : 0.807
Recall : 0.923
F1 : 0.861
###################################
B-CHI
Total assigned by human : 954
Total assigned by algorithm : 1109
Precision : 0.716
Recall : 0.832
F1 : 0.77
###################################
I-CHI
Total assigned by human : 283
Total assigned by algorithm : 313
Precision : 0.546
Recall : 0.604
F1 : 0.574
###################################
B-ADJ
Total assigned by human : 1336
Total assigned by algorithm : 1253
Precision : 0.546
Recall : 0.512
F1 : 0.528
###################################
I-ADJ
Total assigned by human : 588
Total assigned by algorithm : 660
Precision : 0.486
Recall : 0.546
F1 : 0.514
###################################
B-CRO
Total assigned by human : 199
Total assigned by algorithm : 230
Precision : 0.696
Recall : 0.804
F1 : 0.746
###################################
I-CRO
Total assigned by human : 70
Total assigned by algorithm : 98
Precision : 0.643
Recall : 0.9
F1 : 0.75
###################################
B-INC
Total assigned by human : 3074
Total assigned by algorithm : 2391
Precision : 0.698
Recall : 0.543
F1 : 0.61
###################################
I-INC
Total assigned by human : 1431
Total assigned by algorithm : 1176
Precision : 0.675
Recall : 0.555
F1 : 0.609
###################################
O
Total assigned by human : 43226
Total assigned by algorithm : 43180
Precision : 0.994
Recall : 0.993
F1 : 0.994
###################################
```

- In section 4.4 Assessing the model performance against the human reviewed data, the 4th cell provide information for **table 4.**

```
T
Total assigned by human : 796
Total assigned by algorithm : 849
Precision : 0.887
Recall : 0.946
F1 : 0.916
###################################
P
Total assigned by human : 1472
Total assigned by algorithm : 1294
Precision : 0.944
Recall : 0.83
F1 : 0.884
###################################
C
Total assigned by human : 350
Total assigned by algorithm : 260
Precision : 0.931
Recall : 0.691
F1 : 0.793
###################################
A
Total assigned by human : 380
Total assigned by algorithm : 364
Precision : 0.56
Recall : 0.537
F1 : 0.548
###################################
X
```

```
Total assigned by human : 50
Total assigned by algorithm : 58
Precision : 0.655
Recall : 0.76
F1 : 0.704
###################################
I
Total assigned by human : 498
Total assigned by algorithm : 756
Precision : 0.443
Recall : 0.673
F1 : 0.534
###################################
O
Total assigned by human : 9849
Total assigned by algorithm : 9814
Precision : 0.995
Recall : 0.992
F1 : 0.994
###################################
```

- In section 4.4 Assessing the model performance against the human reviewed data, the 7th cell provide a confusion matrix (BERT prediction vs Human labels)

```
knitr::include_graphics("./figs/fig_3.png")
```

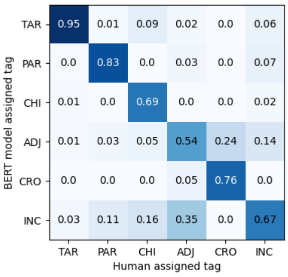


Figure 4: Reproduction of Figure 4