Reproducibility review: "Comparing supervised learning algorithms for Spatial Nominal Entity recognition"

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/suwpj/

To cite this report use

> *Ostermann, F. O., and Nüst, D. (2020, July). Reproducibility review of: Comparing supervised learning algorithms for Spatial Nominal Entity recognition. https://doi.org/10.17605/OSF.IO/SUWPJ*

Reviewed paper

> *Amine Medad, Mauro Gaio, Ludovic Moncla, Sébastien Mustière and Yannick Le Nir: Comparing supervised learning algorithms for Spatial Nominal Entity recognition. AGILE GiScience Ser., 1, 15. https://doi.org/10.5194/agile-giss-1-15-2020, 2020.*

Source code: https://github.com/MedadAmine/Spatial-nominal-entity-recognition

Summary

The authors have done a commendable job at providing all required input data, scripts, and documentation to run the analysis. The reproduction was hindered because differences in computational environment required some initially undocumented adjustments for the libraries used, which have now been documented. It should be noted that the analysis requires substantial downloads, disk space, and processing power to run. Eventually, the reproduction was mostly successful.

Reproducibility reviewer notes

The materials on GitHub have an MIT license.

Data

Original hiking texts: not available, although there is a list of words

Lexicon: FastText freely available online

Corpus: entire corpus not available, although there is a list of words

Samples for analysis available (named corpus), but not documentation as to the meaning

Processing

- uses open source libraries

- Scripts and hyper-parameters are available

- using requirements.txt to install libraries in new virtual environment throws error (incompatible versions), fixed through manual install of libraries

- pre-trained FastText model is massive to download

- example for installation path of model doesn't match load path in scripts

- cudart64 error (ignored) for Tensorflow, depending on GPU

- TreeTaggerError: "Can't locate TreeTagger directory (and no TAGDIR specified)"

We were able to resolve the TreeTagger issue by following the TreeTagger installation instructions (https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/) and downloading the French parameter file. This has now been documented in the repository as well. Afterwards we were able to execute all cells in the provided Jupyter Notebook within a local container (using repo2docker with the --editable option).

Results

The direct link between paper and code/models still has to be inferred. The outputs showed small numerical differences as shown in this commit:

https://github.com/reproducible-agile/Spatial-nominal-entity-recognition/commit/872d110b507e3ba94aa1a23f29fa8539bc9255ff

Some suggestions for further improvements on an already very commendable effort:

- The README should clearly mention the datasource of the download (Facebook AI research?)

- Maybe you could provide a suitable test dataset of a more manageable size, for demonstration and testing; this would even allow to share your workflow as a Binder (see https://mybinder.org)

- Lastly, you could mention your execution times (along with a description of the used hardware)