Review: LDOW2016 manuscript 6

- Permalink: https://gist.github.com/stain/7c74026d8741b4ab74da34b1eca0e9bc
- Title: Evaluating approaches for supervised semantic labeling
- Authors: Natalia Ruemmele, Yuriy Tyshetskiy, Alex Collins
- Call: LDOW2018
- Submitted preprint: https://arxiv.org/abs/1801.09788
- Resource: https://github.com/NICTA/serene-benchmark
- Review by: Stian Soland-Reyes (3/3)
- Recommendation: 1: Weak Accept
- Outcome: TODO

This review is licensed under a Creative Commons Attribution 4.0 International License
http://creativecommons.org/licenses/by/4.0/

Overall evaluation

1: weak accept

This paper compares several machine learning approaches for semantic labelling of relational data sources like CSV files. Compared to earlier approaches this also explores how to add unknown labelling rather than just classification into existing semantic labels.

The authors give great detail on the learning techniques and approaches to sampling and the result achieved when learning, which they claim give reasonable good results even for simple approaches like character frequencies.

The benchmark is available in GitHub as open source, and I was eventually able to install and execute this. However this required a significant amount of effort to install because of the dependency chain. To improve on this, I recommend the authors make their Python libraries and tools available in PyPi for easier installation, as well as provide the Docker image on the Docker Hub. (A Docker Compose file might be needed to combine the different services for the benchmark.)

The source code repositories and datasets has not been made permanently available, e.g. given a DOI with Zenodo's GitHub integration.

For some reason the authors have used datasets of Weapons databases as part of the training data, I find this highly questionable for ethical reasons, and would request that this is not given any prominence in the event of conference presentation of this work.

The paper does not explore if or how any generated Linked Data can be made available on the web, and no Linked Data datasets have been made available. As far as I can tell the main link to the "Web" here has to do with the CSV files previously scraped from websites. As such I am not fully confident in recommending this work as relevant to the LDOW2018 workshop.

Given the above I'm afraid my recommendation is "Weak Accept" - however I believe this work could be valuable to target a venue focusing on benchmarks and forming semantic from unstructured data, such as the ISWC Resources Track.

Reviewer's confidence

4: High
Open Review

I am happy for LDOW to publish my review online.

This open review is also available at the URL https://gist.github.com/stain/7c74026d8741b4ab74da34b1eca0e9bc and have been submitted to the Linked Open Research Cloud.

Transparent Review

I am Stian Soiland-Reyes http://orcid.org/0000-0001-9842-9718

Confidential remarks for the program committee

As the other reviewers indicate their this paper of being relevant for the workshop I will not be blocking its acceptance.