Chris Straschewski

CMS180014

CS 4395.001

ACL Paper Summary

My article was called *Dataset Geography: Mapping Language Data to Language Users*, and had three authors listed: Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos, from the Department of Computer Science of George Mason University, in the US. The main problem introduced by the paper is that with the use of language technologies becoming more and more widespread, language diversity and coverage of NLP systems must be extended to reach everyone. According to the authors, one of the most important factors of how well an NLP system works is data availability. Because of this, they took the approach of studying the “geographical representativeness” of NLP datasets, using metrics like entity recognition and linking systems. Finally, they explore some economic and geographical factors that could further explain the results. With all of this, the main goal is to measure if the NLP datasets in a certain region are up to par and match the needs of the language speakers.

Fahim Faisal’s prior works include discussions of various other NLP applications such as Question Answering systems and Code to Comment translations. Nothing specifically related to this paper, but both related by the strings of NLP. Antonis Anastasopoulos’ include similar topics but over a much wider range. They are clearly quite experienced with NLP applications and research.

A unique contribution of the paper is the presented method of estimating how culturally representative a dataset is by mapping it to a country and visualizing relationships between relevant variables. For example, it was found that dataset representativeness has a strong correlation to a country’s GDP, following by geographical proximity, then population. The paper also chooses a technique for entity linking that removes the need for named entity recognition, and when tested, found that similar dataset mappings were found anyway. These dataset mappings are found throughout the paper, providing visualizations of the results of the work.

Overall, the conclusion of the paper was that the authors provided a new way for effectively visualizing the cultural representativeness of NLP datasets, but that there is still more work to be done, as the tool can still be improved upon. Author Fahim Faisal received 59 citations on google scholar, and Antonis Anastasopoulos received 2316. All their work is important because they have done significant amounts of research into NLP applications, and the only way to improve upon things like NLP is to do just that. Research, find strengths and weaknesses, and work to improve upon them.