Designing computational systems to support humanities and social sciences research

Eetu Mäkelä
University of Helsinki, Finland
P.O. Box 24, 00014
eetu.makela@helsinki.fi

Abstract

From the viewpoint of the humanities and social sciences, collaborations with computer scientists often fail to deliver. In my research group, we have tried to understand why this is, and what to do about it. In this talk, I will discuss three key elements that we have discovered:

Often, datasets in the humanities and social sciences are not neatly representative of the object of interest. Systems need to provide ways in which to evaluate and counter the biases, confounders and noise in the data. Often, there is also a large gap between what is in the data, and what would be of interest. This gap needs to be bridged using algorithms, but care must be given that a) what the algorithm produces actually matches the interest and b) that its application does not introduce bias of its own (also interestingly, algorithm performance metrics of interest here often differ from those generally used in NLP/computer science). On a process level, collaboration between researchers from different disciplines is hard due to discrepancies in expectations relating to all facets of research, from research questions through methodology to the publication of results. Projects and systems need to acknowledge this, and be designed to facilitate iterative movement in the right direction.

Bio

Eetu Mäkelä is an associate professor in Human Sciences–Computing Interaction at the University of Helsinki, and a docent (adjunct professor) in computer science at Aalto University. At the Helsinki Centre for Digital Humanities, he leads a research group that seeks to figure out the technological, processual and theoretical underpinnings of successful computational research in the humanities and social sciences.

Additionally, he serves as a technological director at the DARIAH-FI infrastructure for computational humanities and is one of three research programme directors in the datafication research initiative of the Helsinki Institute for Social Sciences and Humanities. For his work, he has obtained a total of 19 awards, including multiple best paper awards in conferences and journals, as well as multiple open data and open science awards. He also has a proven track record in creating systems fit for continued use by their audience.

Large-scale language models: challenges and perspective

Benoît Sagot
Inria Paris (équipe ALMAnaCH)
2 rue Simone Iff CS 42112
75589 Paris Cedex 12, France
benoit.sagot@inria.fr

Abstract

The emergence of large-scale neural language models in Natural Language Processing (NLP) research and applications has improved the state of the art in most NLP tasks. However, training such models requires enormous computational resources and training data. The characteristics of the training data has an impact on the behaviour of the models trained on it, depending for instance on the data's homogeneity and size. In this talk, I will speak about how we developed the large-scale multilingual OSCAR corpus. I will describe the lessons we learned while training the French language model CamemBERT, the first large-scale monolingual model for a language other than English, especially in terms of the influence of size and heterogeneity of the training corpus. I will also sketch out a few research questions related to biases in large-scale language models, with a focus on the impact of tokenisation and language imbalance, in the context of the BigScience initiative. I will conclude with my thoughts on the future of language models and their impact on NLP and other data processing fields (speech, vision).

Bio

Benoît Sagot, Directeur de Recherches (Senior Researcher) at Inria, is the head of the Inria project-team ALMAnaCH in Paris, France. A specialist in natural language processing (NLP) and computational linguistics, his research focuses on language modelling, language resource development, machine translation, text simplification, part-of-speech tagging and parsing, computational morphology and, more recently, digital humanities (computational historical linguistics and historical language processing). He has been the PI or co-PI of a number of national and international projects, and is the holder of a chair in the PRAIRIE institute dedicated to research in artificial intelligence. He is also the co-founder of two start-ups where he uses his expertise in NLP and data mining for the automatic analysis of employee survey results.