There is no data like well-chosen data: Active Learning for humanities data annotation

Folgert Karsdorp*, Peter van Kranenburg*
*Meertens Institute
{folgert.karsdorp,peter.van.kranenburg}@meertens.knaw.nl

Manually annotating data collections often is a highly costly endeavor. This holds especially true in the Humanities where scholars deal with high dimensional, complex data. Automatic systems from Machine Learning or Information Retrieval can assist in such annotation projects. Typically these systems need to be trained on a set of annotated data on the basis of which they 'learn' how to annotate new, unseen data. Automatically obtained annotations can then be evaluated by human annotators. This co-operation between human and machine greatly reduces the intense cost of manually annotating data. A common characteristic of these automatic systems, however, is that they require large amounts of training data to be put to action reliably. A famous quote by Mercer's reads "There is no data like more data". It addresses the issue that in many Machine Learning problems more data is often considered to be more important than better algorithms. Notwithstanding this adage, obtaining such large amounts of annotated data is often not feasible, at least, not without the help of a machine. To reduce the cost of a brute-force data collection struggle, we propose an iterative data annotation pipeline – inspired by Active Learning techniques – which tries to optimize the learning curve of a learning system. The key feature of this pipeline is that it makes a ranking of the order in which new items should be annotated to more efficiently arrive at a well-preforming system. Most modern Machine Learning systems (e.g. Support Vector Machines) utilize some kind of decision function which is used to classify objects into different classes. The values of this decision function can be seen as representations of the certainty with which the system made its classifications. We show that by focussing the annotation endeavor first and foremost on those cases where the learning system exhibits the highest classification *uncertainty*, we are able to develop a better learning system with less annotation effort. To establish empirical ground for this hypothesis, we evaluate the impact of the annotation pipeline on the learning curve using a range of established machine learning data sets. We then proceed with a more in-depth evaluation of two annotation experiments in two different subfields of computational humanities. In the first the goal is to annotate a large collection of folktales with motifs of magical transformation. The second is aimed at recognizing melodic phrase endings (melodic cadences) in folk songs.