Bootstrapping Clinical Concept Extraction with Self-training

Nyein Pyae Pyae Khin
nyeinpyaepyaekhin@gmail.com
University of Computer Studies, Mandalay, UCSM

Khin Thida Lynn
khinthidalinn@ucsm.edu.mm
University of Computer Studies, Mandalay, UCSM

In the clinical domain, annotated clinical records are not only expensive but also often unavailable for research due to patient privacy and confidentiality requirements. The challenge is how to train effective clinical concept extraction system especially with small amount of training data. To address the limited supervision problem of insufficient labeled training examples, self-training style semi-supervised bootstrapping approach to concept extraction system is proposed. In self-training a classifier is trained from an initially small amount of human annotated data, and then used to label unlabeled data. The machine-labeled data is then added to the original data set, and the classifier is retrained iteratively. For labeling clinical concepts, Conditional Random Fields (CRF) is chosen due to its promising performance in many sequence labeling tasks.