Data-driven discovery of medical terms from Chinese electronic health records

Sheng Yu

Tsinghua University
E-mail: syu@tsinghua.edu.cn

Abstract: A comprehensive medical terminology is the basis for mining electronic health records (EHR) and a key infrastructure for medical big data analysis. Chinese medical terminology development is extremely lacked behind compared to English, and severely hampers the development of medical artificial intelligence in China. We propose a method that identifies medical terms from the EHR for the automatic construction of a medical terminology. We treat a sentence as an undirected graph, whose nodes are the characters in the sentences, and whose edge weights represent the connection strength computed with corpus statistics – larger weighs indicate the associated characters are more likely to be in the same term/word. The word segmentation is then achieved with spectral graph partition in an unsupervised manner. After segmentation, a Bi-LSTM classifier is applied to remove incorrectly segmented words and nonmedical words/terms from the output.