

Multi-class Biomedical Term Recognition and Classification Using Search Engines

Raul Rodriguez-Esteban
Columbia University, USA

Abstract:

The talk will have two parts:

(1) Automatic curation of biomedical text mined data. Biomedical text mined data has a degree of quality sometimes unsuitable for real-world applications. Manual curation of data improves quality but at high cost. We introduce a new method to automatic curation that is based in machine learning techniques and allows us to curate large repositories.

(2) Multi-class biomedical term identification using search engines Multi-class biomedical term identification lags behind in terms of performance. Ideally, there is a large a number of term classes that are of potential interest in biomedical text and are not being extracted. To increase the number of term classes identified, we have devised a two-layer process that separates term recognition from term classification. The first layer identifies any potential terms using a grammatical model and the second layer performs the classification of terms of interest into a large set of classes defined by examples. To improve both recognition and classification we introduce the use of features generated from searches in large indexed biomedical corpora.

Short Bio:

Raul Rodriguez-Esteban has just defended his Ph.D. at the Department of Electrical Engineering of Columbia University. He is a member of the National Center for Biomedical Computing (MAGNet) and the Center for Computational Biology and Bioinformatics at Columbia University.