Clinical effect extraction from eastern medicine database

Jaehyun Lee[1], Yeeok Kang[1, 2], Doheon Lee[1,*]

[1]Korea Advanced Institute of Science and Technology, Department of Bio and Brain Engineering, Korea
[2]SD Genomics Co, Korea

Abstract

Traditional Chinese medicine (TCM) is empirically accumulated knowledge for over 2,000 years. Traditional Chinese medicines integrated database (TCMID) is a large-scale TCM database with the unstructured plain text describing the functions of herbs and formulae. To analyze the high dimensional data using computational approaches for clinical effect extraction, the plain text should be structured by entity recognition (clinical entities and effective entities) and clinical/effective entity pairing. This paper presents a pilot study and its result to extract specific and atomic clinical effects from the plain text based on the machine-learning approach. The main task was divided into two independent steps which were defined as supervised learning problems: effective entity recognition and clinical/effective entity pairing. 100 herb entries and their functional descriptions from TCMID were randomly selected to generate the training corpus by manually tagging clinical effects. For the preprocessing step, MetaMap and BLLIP parser were utilized. MetaMap which is a tool for recognizing UMLS concepts was to recognize clinical entities according to the corresponding semantic types. And then, BLLIP parser identified the deep-parsed structures and extracted syntactic features for support vector machine (SVM) modelling. Based on the feature set, the two SVM classifiers were learned for effective entity recognition and clinical/effective entity pairing, respectively. The proposed pipeline achieved an F-score of 88.97% on the eventual task. Therefore, clinical effect extraction to organize plain text in TCMID would promise the time- and cost-saving approach for drug developers to analyze TMC databases in an automated manner.

Biography

Jaehyun Lee received M.S. degree in Bio and Brain Engineering form Korea Advanced Institute of Science and Technology (KAIST), Korea in 2014 and is currently in the Ph.D. course in the aforementioned affiliation. His research interests include bioinformatics, text-mining, and machine learning.

jaeh@kaist.ac.kr