Medical Subdomain Classification Using Unstructured Clinical Documents and Machine Learning-Based Natural Language Processing Approach

Wei-Hung Weng, MD1,2, Kavishwar B. Wagholikar, MBBS, PhD3,1, and Henry C. Chueh, MD, MS2,3
1 Department of Biomedical Informatics, Harvard Medical School, Boston, MA
2 Laboratory of Computer Science, Department of Medicine, Department of Neurology, Massachusetts General Hospital, Boston, MA

Abstract

The medical subdomain of a clinical document is a useful metadata for developing machine learning models. To classify the medical subdomain of a document accurately, we have constructed an automated machine learning-based natural language processing (NLPI) pipeline and developed accurate machine subdomain classifiers based on the content of the document. We constructed the pipeline using cTAKES, the UMLS Metathesaurus and Semantic Network, and learning algorithms to extract features from two datasets — documents from the iDASH repository (n = 431) and clinical notes from Massachusetts General Hospital (MGH) (n = 91,237), and built medical subdomain classifiers with different combinations of clinical feature representations and learning algorithms. The performance of classifiers and model portability across two datasets was evaluated. The linear support vector machine trained machine subdomain classifier using hybrid bag-of-words and clinically relevant UMLS concepts as the feature representation, with term frequency-inverse document frequency weighting, outperformed other classifiers on iDASH and MGH datasets with F1 scores of 0.932 and 0.934, and areas under ROC curve (AUC) of 0.957 and 0.964, respectively. Classifiers for half of medical subdomains were found to be portable across two datasets at the threshold of F1 score of 0.7. Our study shows that the machine learning-based NLP approach, it is possible to develop medical subdomain classiﬁers with sufﬁcient performance for clinical applications. Portable medical subdomain classiﬁers may also be used across different datasets from different institutions.

Motivation

• The medical subdomain, such as cardiology, genoproteomics and neurology, is metadata of a document that is necessary to enhance the effectiveness of clinical machine learning-based NLP models by considering specialty-associated conditions
• Unstructured clinical documents have been regarded as a powerful resource to solve different clinical questions by providing detailed patient conditions, the thinking process of clinical reasoning, and clinical inference, which usually cannot be obtained from the other components of the electronic health record (EHR) system
• Automated clinical document classification is an essential component of clinical predictive analytics that can extract knowledge and categorize documents into predefined document-level thematic labels

Current / Proposed Approach

• Automated document classification task was usually performed via rule-based knowledge engineering, by manually implementing a set of expert intelligence rules
• Machine learning-based NLP approach — NLP-derived feature representation and supervised learning algorithm using pre-classiﬁed documents
• MEDLINE documents
• Hybrid word and phrase representation with a support vector machine (SVM) algorithm
• The Medical Subject Headings (MeSH) ontology as a feature representation with a maximum entropy algorithm
• Chest radiograph reports
• Medical Language Extraction and Encoding System (MedLEE) with UMLS Metathesaurus to identify medical concepts and classify into six clinical conditions

Research Objectives

• Developed an automated machine learning-based NLP pipeline to build subdomain classifiers
• Examined the performance between the classifiers using different clinical feature representation methods, weighting strategies, and supervised learning algorithms
• Investigated the portability of classifiers across two real-world clinical datasets

Conclusion

The purpose of the study is to classify the medical subdomain of an unstructured or semi-structured clinical document accurately, and we proved that the machine learning-based classification model could be an optimal solution for building portable medical subdomain classifiers. Using two sets of clinical notes, we found that the selection of the classifier building combination of the clinical feature representation and supervised learning algorithm is important to yield a better-performed and portable medical subdomain classifier. We plan to integrate the information of both medical subdomain and clinical expert to build hierarchical models for consolidating our methods, and may adopt techniques of domain adaptation and transfer learning to improve the performance of model portability and construct a well-generalizable solution.

NLP and Machine Learning

• Using cTAKES, UMLS Metathesaurus and Semantic Network. [1-3]
• 98 classifiers for each dataset
• Seven feature representation methods
• Hybrid word and phrase concepts / UMLS concepts / Concepts restricted by semantic groups or types / words-concepts combinations
• Two vector representation methods
• Frequency count / tf-idf
• Supervised learning algorithms
• Naïve Bayes, multinomial logit regression (1.1, 1.2 penalization), linear SVM
• Adaptive boosting
• Accuracy, precision, recall, F1, AUC

Materials and Methods

Data Source

• 431 clinical documents from the iDASH repository + 91237 MGH clinical notes from Partners HealthCare RPDR (IRB#16E000011)

References