Published online: 2019-08-16

Appendix: Content Summaries of Best Papers for the Health Information Management Section of the 2019 IMIA Yearbook

Atutxa A, Pérez A, Casillas A

Machine Learning Approaches on diagnostic term encoding with the ICD for clinical documentation

IEEE J Biomed Health Inform 2018;22(4):1323-9

This study focuses on data mining applied to unstructured clinical text in electronic health records (EHRs). The authors tried to improve standard machine learning techniques. They believe that clinical text mining can efficiently leverage the encoding process and they sought to develop computer-assisted classification tools and applications to help coding experts. The paper focuses on developing automatic techniques to encode diagnostic terms (DTs), focusing on Spanish language EHRs and publicly available resources for Spanish clinical text processing and mining.

In the study, records were encoded manually by experts using the International Classification of Diseases (ICD). The dataset consisted of spontaneous DTs extracted from over 9,000 EHRs, with 1,500 possible ICD codes. The proposed system was able to select the correct ICD code with 92% precision for the main disease (primary class) and 88% precision for the main disease together with the non-essential modifiers (fully specified class). The authors note that the methodology is simple and portable with potential applicability for documentation and pharmaco-surveillance. In a pilot study using a small sample of records, experts from public hospitals reported an accuracy of 91.2%. The authors have made the software publicly available so that the techniques and approach can be used by a broader audience of both clinicians and researchers.

Cui L, Xie X, Shen Z

Prediction task guided representation learning of medical codes in EHR

J Biomed Inform 2018;84:1-10

The authors review applications using machine learning models for predictive analytics in electronic health records (EHRs). Machine learning has been used to improve the quality and efficiency of services. Developing machine learning models requires converting medical codes representing diagnoses and procedures to feature vectors. The authors recognize the importance of vector representations on the performance of machine learning models. They sought to address shortcomings of previous efforts using representation learning methods from Natural Language Processing (NLP) to learn vector representations of medical codes. As stated by the authors, "the objective of the study was to develop a representation learning model which can learn vector representations of medical codes that have strong predictive capability for various prediction tasks and required relatively small amounts of training data". The researchers used a dataset that contained 750,000 cases and represented three years of records from five hospitals in China. Researchers developed a new method that they called "Prediction Task Guided Health Record Aggregation (PTGHRA)" which aggregates health records guided by prediction tasks, to construct a training corpus for various representation learning models. PTGHRA uses representation learning methods to map medical codes to continuous vectors and combined medical code vectors with other information in health record to form feature vectors for prediction tasks. Authors focused on the prediction of cost and length of stay (LoS). Compared with unsupervised approaches, representation learning models integrated with PTGHRA yielded a significant improvement in predictive capability of generated medical code vectors. For training set sizes smaller than 20,000 records, PTGHRA achieved up to 32% accuracy improvement.

Li F, Liu W, Yu H

Extraction of information related to adverse drug events from electronic health record notes: design of an end-to-end model based on deep learning

JMIR Med Inform 2018;6(4):e12159

This paper focuses on how pharmacovigilance and drug-safety surveillance are crucial for monitoring adverse drug events (ADEs) and notes challenges (such as under-reporting) with several existing ADE-reporting systems. The main purpose of the study was to develop a deep learning model focusing on identifying ADEs, medications, and indications. A secondary purpose was to improve the deep learning model. The authors used the Medication, Indication, and Adverse Drug Events (MADE) 1.0 challenge to develop both training and testing datasets. MADE contains data (1,089 EHR notes) from cancer patients and includes nine entity types including Medication, Indication, and ADE. It also includes seven types of relations between these entities. The training data included 876 and the testing dataset used for the remaining 213 notes. To extract information from the dataset, the authors used a deep-learning model applying bidirectional long short-term memory (BiLSTM) conditional random field network to recognize entities and a BiLSTM-Attention network to extract relations. They enhanced their deep learning model with three multitask learning (MTL) methods (hard parameter sharing, parameter regularization, and task relation learning). The authors used the results of the second step of the process of extracting ADE information (relation extraction) to compare all models. They used micro-averaged precision, recall, and F1 as evaluation metrics. The authors compared their model with the top three systems in the MADE 1.0 challenge. Their model achieved state-of-the-art results (F1=65.9%). The model using hard parameter sharing further improved the F1 by 0.8%, boosting the F1 to 66.7%. The authors concluded that the performance of ADE-related information extraction can be improved by employing deep learning and MTL models. They also felt that the methods, data, and other factors influence the effectiveness of MTL models. The authors believe that this study could be useful for further natural language processing and machine learning research for detection of adverse drug events in clinical notes. Their annotated Dataset (Medication, Indication,

and Adverse Drug Events (MADE)) -- will be publicly available to support research on extraction of ADE-related information.

Qiu JX, Yoon H-J, Fearn PA, Tourassi GD

Deep learning for automated extraction of primary sites from cancer pathology reports

IEEE J Biomed Health Inform 2018;22(1):244-51

The paper focuses on using data extracted from pathology reports to populate cancer registries. The authors explore using a machine learning approach to make the labor-intensive manual process of information extraction and coding easier. In this study, researchers investigated deep learning and a convolutional neural network (CNN) for extracting ICD-O-3 (International Classification of Diseases for Oncology) topographic codes from breast and lung cancer pathology reports. They compared a CNN with a term frequency vector approach, using 942 de-identified pathology reports matched to 12 ICD-O-3 topography codes corresponding to seven breast and five lung primary sites. They compared the results to reviews by human (cancer registry) subject matter experts. Pathology reports were provided from five different Surveillance, Epidemiology and End Results (SEER) cancer registries. Cancer registry experts manually annotated the pathology reports and their annotations served as the gold standard. Researchers observed that when class labels were well populated, deep learning models (e.g., CNN) outperformed the conventional approaches (term frequency vector approaches) in the studies looking at class prevalence (micro- and macro-F score increases of up to 0.132 and 0.226). The increase in performance of deep learning methods from transfer learning was less strong and depended on the CNN method and cancer site. The authors believe that their results demonstrate the potential of deep learning for automated abstraction of relevant information for cancer registries from pathology reports.