Distant Supervision in Clinical Information Retrieval

Xing Wei, MSc[1], Carsten Eickhoff, PhD[2]
[1]ETH Zurich, Zurich, Switzerland; [2]Brown University, Providence, RI

Abstract

Neural network frameworks have recently shown highly effective at a wide range of tasks ranging from radiography interpretation via data-driven diagnostics to clinical decision support. This often superior performance comes at the cost of increased training data requirements that cannot be satisfied in every given institution or scenario. This paper presents a distant supervision approach that alleviates the need for scarce in-domain data by relying on a related, resource-rich, task for training.

Materials & Methods

This study presents an end-to-end neural network clinical decision support system that recommends relevant literature for individual patients based on their unstructured electronic health records (EHR). As a benchmarking testbed for this application the TREC Clinical Decision Support track [1] provides 30 manually annotated free-text EHRs along with lists of Medline articles of confirmed relevance to the patient at hand. Highly parametric neural network methods have been repeatedly shown to unfold their full predictive potential only when supplied with sufficient amounts of training examples. This observation lets us hypothesize that neural text matching architectures will struggle to meet or surpass state-of-the-art performance with such limited training data. To overcome this limitation, we turn to a secondary, related task for which training data is abundant and on which models can be pre-trained.

Concretely, we design a convolutional neural network classification scheme [2] that learns to associate 43k MIMIC-III patients [3] with their respective diagnoses. Trained in this resource-abundant domain, the feature representation component is subsequently integrated into an end-to-end text matching architecture [4] to deliver clinical decision support in the original, resource-impoverished domain.

Results

Our experiments highlight two key findings: **a)** As hypothesized, exclusively locally-trained models fail to meet the quality standards of state-of-the-art term-based models. **b)** Using distant supervision on a resource-rich related task introduces statistically significant and substantial improvements of up to 40% relative performance for the original task.

Discussion

Clinical data science and machine learning are frequently faced with scenarios in which data is prohibitively "small", preventing application of more effective methods. The distant supervision paradigm presented in this study has potential for broad application in a wide number of settings that suffer from low incidence rates.

References

1. Roberts K, Simpson MS, Voorhees EM, Hersh WR. Overview of the TREC 2015 Clinical Decision Support Track. In: TREC; 2015. .

2. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports. 2016;6:26094.

3. Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3.

4. Guo J, Fan Y, Ai Q, Croft WB. A Deep Relevance Matching Model for Ad-hoc Retrieval. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16. New York, NY, USA: ACM; 2016. p. 55–64. Available from: `http://doi.acm.org/10.1145/2983323.2983769`.