International Journal of Medical Informatics 105 (2017) 121–129

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Automated problem list generation and physicians perspective from a pilot study

Murthy V. Devarakonda[a,*], Neil Mehta[b], Ching-Huei Tsou[a], Jennifer J. Liang[a], Amy S. Nowacki[b], John Eric Jelovsek[b]

[a] *IBM Research, USA*
[b] *Cleveland Clinic, USA*

ARTICLE INFO

Keywords:
Electronic health records
Longitudinal patient records
Problem list
Machine learning
Natural language processing
IBM Watson

ABSTRACT

Objective: An accurate, comprehensive and up-to-date problem list can help clinicians provide patient-centered care. Unfortunately, problem lists created and maintained in electronic health records by providers tend to be inaccurate, duplicative and out of date. With advances in machine learning and natural language processing, it is possible to automatically generate a problem list from the data in the EHR and keep it current. In this paper, we describe an automated problem list generation method and report on insights from a pilot study of physicians' assessment of the generated problem lists compared to existing providers-curated problem lists in an institution's EHR system.

Materials and methods: The natural language processing and machine learning-based Watson[1] method models clinical thinking in identifying a patient's problem list using clinical notes and structured data. This pilot study assessed the Watson method and included 15 randomly selected, de-identified patient records from a large healthcare system that were each planned to be reviewed by at least two internal medicine physicians. The physicians created their own problem lists, and then evaluated the overall usefulness of their own problem lists (P), Watson generated problem lists (W), and the existing EHR problem lists (E) on a 10-point scale. The primary outcome was pairwise comparisons of P, W, and E.

Results: Six out of the 10 invited physicians completed 27 assessments of P, W, and E, and in process evaluated 732 Watson generated problems and 444 problems in the EHR system. As expected, physicians rated their own lists, P, highest. However, W was rated higher than E. Among 89% of assessments, Watson identified at least one important problem that physicians missed.

Conclusion: Cognitive computing systems like this Watson system hold the potential for accurate, problem-list-centered summarization of patient records, potentially leading to increased efficiency, better clinical decision support, and improved quality of patient care.

1. Background and significance

Electronic Health Records (EHRs[2]) are expected to improve patient outcomes by providing the most important patient information in a single location [1]. A common way to provide the holistic information is in the form of a patient-centered problem list [2], by itself or as part of a summary [3,4]. Ideally, the problem list would include all clinically significant issues that a care provider should consider when managing a patient. It should be accurate, inclusive, and up to date [5,6]. At the same time, it should not contain redundant or irrelevant items that distract the clinician, reduce efficiency, or lead to inappropriate actions [7–9].

Existing EHR systems provide problem lists that are populated by care providers, but they tend to suffer from the problems outlined. A patient's problem list in an EHR system requires the clinician to routinely maintain and update it with each encounter. Unfortunately, this is haphazardly performed, leading to inaccurate, incomplete and duplicative lists.

* Corresponding author.
E-mail address: mvd@acm.org (M.V. Devarakonda).

[1] ***Watson***, mentioned here, refers to ***new methods*** developed ***for patient record text analytics***, including the automated problem list generation, based on the core Watson text processing tools for sentence segmentation, parsing, and named entity linking.

[2] In this article, we use the terms ***EHR*** and ***EHR system*** to mean commercial and non-commercial electronic health record systems, and we use the term ***patient record*** to mean all the patient data, including clinical notes, reports, medications ordered, procedures, labs, and demographics.

http://dx.doi.org/10.1016/j.ijmedinf.2017.05.015
Received 28 December 2016; Received in revised form 21 May 2017; Accepted 23 May 2017

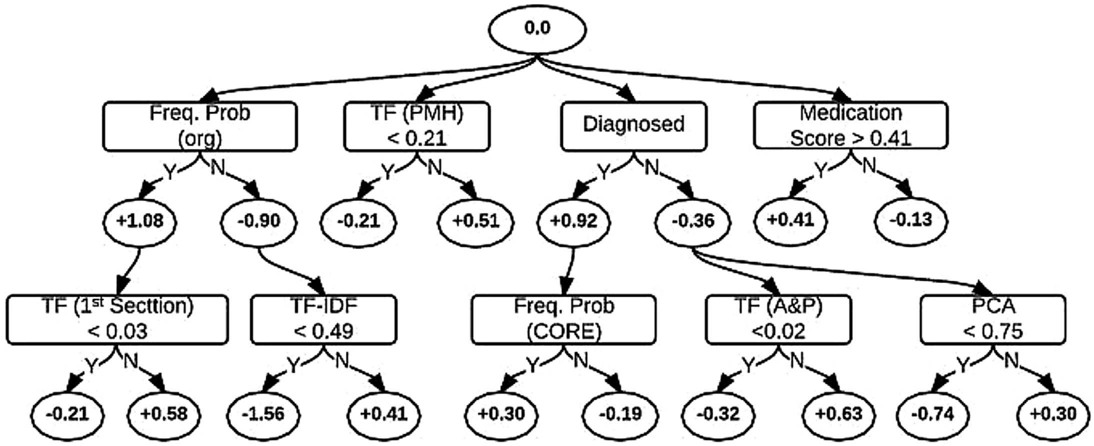

Fig. 1. The first two levels of the Watson problem list ADT model. Rectangular nodes (splitting nodes) represent conditions while oval nodes (prediction nodes) contain numeric scores that are used to make a prediction. The full model is 7 levels deep and contains 50 splitting nodes.

Efforts to automate the creation of the problem list based on EHR data have been made, but were limited to 80 prespecified problems and utilized techniques that may not scale to all possible problems [10–12]. With the advent of advanced natural language processing (NLP) and machine learning techniques [13–15], it is possible for automatic creation and updating of a patients' problem list based on the data in the EHR. We have developed a system based on these technologies, present its key features and report a pilot study of its usefulness from a physician's perspective. The study was conducted in USA.

2. Method: automated problem list generation

The Centers for Medicare and Medicaid Services [16] of the USA defines a patient's problem list as, "a list of current and active diagnoses as well as past diagnoses relevant to the current care of the patient" for the federal meaningful use program. We made it more concrete by assuming the context to be a comprehensive health assessment and by including clinically relevant past procedures in the problem list.

2.1. Framing the machine learning task

We framed the problem list generation as a supervised binary classification task, which decided if a candidate problem is a true problem or not. Candidate problems were identified from the textual narrative of clinical notes using NLP (details in next subsection). Several features were extracted for each candidate, and a trained model was applied to the feature values to determine a confidence score for each candidate problem. If the score was above a learned threshold, then the candidate problem was considered a true problem. We engineered features based on clinical, lexical, structural, temporal, and epidemiological aspects of candidate problems, and the features were extracted using NLP from narratives in all types of clinical notes and reports, and from structured parts of a patient record. The machine learning model was trained on a gold standard that was manually created by medical experts (who were not participants in the pilot study) using 399 of the total 996 de-identified patient records obtained from Cleveland Clinic (Cleveland, Ohio, USA) under an Institutional Review Board approval. The following paragraphs provide details of this overview, starting with the step of candidate problems identification.

2.2. Candidate problems identification

Clinical notes were first preprocessed to detect and link medical concepts to the Unified Medical Language System (UMLS®) dictionary entries [17] using Watson NLP components, which provided a function similar to cTAKES [18], but with additional refinements (that are unimportant for the discussion here). Out of these concepts, the subset belonging to the semantic groups *Disorders, Procedures, Physiology,* and *Living Beings* were considered as candidate problems, so long as these concepts had a mapping in the CORE (Clinical Observations Recordings and Encoding) subset of SNOMED CT (Systematized Nomenclature of Medicine − Clinical Terms) [19]. The CORE subset represents a candidate problem space of more than 6,000 problems. Therefore, the candidate problem list would be typically very large (about 15–1) compared to a patient's actual problems. The goal of the machine learning task was to reduce this large set to the actual problem list.

2.3. Features

We engineered a set of features for the machine learning model which would reduce the candidate problems to the actual problem list. At the time of writing this article, the model used 211 features: 18 multi-valued categorical features (e.g. ICD9 category), 28 binary features (e.g. did the candidate problem appear in a clinical note in the last 3 months?), 66 numerical features (e.g. term frequency), and 100 resulting from normalization of the numerical features in various ways (e.g. normalized to the length of the record). For each candidate problem, the features were extracted from the passages (in the clinical notes) where the candidate problem was mentioned, its surrounding context, as well as from the structured data elements related to the problem in the patient record (e.g. diagnosis codes, medications, laboratory test results, and procedures). Details of the features were described in [20]. However, later in this article, we will further discuss some of the features which show how our model represents clinical thinking of medical experts, and therefore played an important role in our trained model.

2.4. Machine learning model

We used the Alternating Decision Tree (ADT) [21], a boosting-based discriminative classifier, to model our classification task due to its accuracy and interpretability. ADT alternates between two types of nodes: a prediction node and a splitter node, represented by an oval and a rectangle in Fig. 1, respectively. During each training step, all features are considered and the best split (i.e., minimizing the training error) is added to the tree. ADT can be viewed as a collection of rules, which predict whether a candidate problem is a true problem or not by adding up the scores of prediction nodes in all valid paths. For example, referring to the ADT in Fig. 1, if a candidate problem is *diagnosed* by a physician and is in the *frequent problems of CORE* then one path for the candidate yields a score of (0.0 + 0.92 + 0.30 =) 1.22. If the aggregate