Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A pilot study of a heuristic algorithm for novel template identification from VA electronic medical record text

Andrew M. Redd (PhD.)[a,b,*], Adi V. Gundlapalli (MD. PhD.)[a,b,c], Guy Divita (PhD.)[a,c], Marjorie E. Carter (MSPH)[a], Le-Thuy Tran (PhD.)[a,c], Matthew H. Samore (MD.)[a,b,c]

[a] VA Salt Lake City Health Care System, Salt Lake City, UT, United States
[b] Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, United States
[c] Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, United States

ARTICLE INFO

Article history:
Received 26 February 2016
Revised 19 July 2016
Accepted 20 July 2016
Available online xxxx

Keywords:
Templates
Natural language processing
Information extraction
Text mining

ABSTRACT

Rationale: Templates in text notes pose challenges for automated information extraction algorithms. We propose a method that identifies novel templates in plain text medical notes. The identification can then be used to either include or exclude templates when processing notes for information extraction.
Methods: The two-module method is based on the framework of information foraging and addresses the hypothesis that documents containing templates and the templates within those documents can be identified by common features. The first module takes documents from the corpus and groups those with common templates. This is accomplished through a binned word count hierarchical clustering algorithm. The second module extracts the templates. It uses the groupings and performs a longest common subsequence (LCS) algorithm to obtain the constituent parts of the templates. The method was developed and tested on a random document corpus of 750 notes derived from a large database of US Department of Veterans Affairs (VA) electronic medical notes.
Results: The grouping module, using hierarchical clustering, identified 23 groups with 3 documents or more, consisting of 120 documents from the 750 documents in our test corpus. Of these, 18 groups had at least one common template that was present in all documents in the group for a positive predictive value of 78%. The LCS extraction module performed with 100% positive predictive value, 94% sensitivity, and 83% negative predictive value. The human review determined that in 4 groups the template covered the entire document, with the remaining 14 groups containing a common section template. Among documents with templates, the number of templates per document ranged from 1 to 14. The mean and median number of templates per group was 5.9 and 5, respectively.
Discussion: The grouping method was successful in finding like documents containing templates. Of the groups of documents containing templates, the LCS module was successful in deciphering text belonging to the template and text that was extraneous. Major obstacles to improved performance included documents composed of multiple templates, templates that included other templates embedded within them, and variants of templates. We demonstrate proof of concept of the grouping and extraction method of identifying templates in electronic medical records in this pilot study and propose methods to improve performance and scaling up.

Published by Elsevier Inc.

1. Introduction

Extracting relevant information from the electronic medical record for clinical, operational, and research purposes is an important and growing field. In the age of big data analytics, it is important to make use of all available information from the patient's medical record, including being able to extract relevant information from the free text of electronic medical notes. While great strides have been made using natural language processing, there remain significant logistical barriers to efficiently processing large corpora of electronic medical notes and generating data that could be used for patient care and other purposes.

Templates and boiler-plated sections, whose origins come from injections of semi-structured text from electronic patient record system pull down menus, are an idiosyncrasy of clinical text in this

[*] Corresponding author at: VA Salt Lake City Health Care System, Salt Lake City, UT, United States.

E-mail addresses: andrew.redd@hsc.utah.edu (A.M. Redd), adi.gundlapalli@hsc.utah.edu (A.V. Gundlapalli).

http://dx.doi.org/10.1016/j.jbi.2016.07.019
1532-0464/Published by Elsevier Inc.

age of electronic medical records (EMR). Instances of these forms need to be treated differently than typical prose. These are often semi-structured check box formats where the provider checks only those relevant affirmative statements and leaves the rest unanswered. Responses contained in the unanswered questions, in the form of numbers, words, or even single characters, are fallaciously counted as positive mentions in traditional natural language processing (NLP) systems [7].

Templates appear in a variety of formats and are used to capture information on the patient. These may be in the form of responses to questions regarding their current symptoms, past medical history, structured questionnaires to score for common conditions such as depression, or a checklist prior to a procedure or surgery. The semantics of each are as idiosyncratic as the variety of formatting found in modern electronic medical record systems, both commercial and home-grown. For example, clinical reminders, when injected into records, pose a particularly vexing problem, because they often assert that the provider talked to the patient about a specific condition. Traditional information extraction systems, not knowing the deeper context of clinical reminders, erroneously pick up those reminders as asserted conditions. In this paper we focus on boiler-plate and copy-paste-edit style templates.

Comparing text to the template from which it originated allows for identification of the information that a clinician inserted or deleted. Since it required intervention and more work for a clinician to type in or remove information, we can place a higher value on that information. Likewise, simply because a phrase is present in a document is not necessarily an indication of important information if the text is part of a template.

Our work on information extraction using US Department of Veterans Affairs (VA) medical record text has revealed that positively asserted concepts of interest are often present in templated text [7]. Thus, rather than exclude these sections from our information extraction pipelines, we have made attempts to develop methods to address the problem of extracting relevant concepts from templated text. We have worked on several domains of interest with regard to the health and health care of Veterans, including homelessness [13,9].

Template identification will provide improved association between answers and text, improving overall information extraction. A straight-forward mechanism to account for such templated sections is to accumulate a database from EMR systems as the basis for searching for templates in clinical notes. While such a database does exist, it is incomplete, does not cover all years, does not cover all of the EMR, and does not account for local, end-user created templates and boiler-plates from copying and pasting from existing medical records. While we have been privy to viewing a VA database of over 40,000 boiler-plate templates from VA records from 2013, the 2015 version contains over 190,000 forms. However, we continue to find instances not found in these resources.

Finding boiler-plate text within clinical text is similar to finding plagiarized content within student papers [11]. It also shares components of the task of finding novel content within newspaper articles [1]. Both tasks have well established techniques [3] to efficiently find what they are looking for. This paper outlines techniques that have their origins and foundations in these techniques.

In this paper we present a pilot of a heuristic algorithm for identification of common templates from a corpus of documents originally presented in [8]. The 750 documents come from 141 different standardized note titles, the most common types being: primary care, primary care outpatient, nursing, mental health, and discharge summary.

We make no assumption of knowledge of the corpus or the templates contained within the corpus, aside from the assumption that the corpus contains multiple instances of documents generated from the same template, a condition that will be met in most all NLP information extraction projects. The framework of information foraging (IF) [12] provides the theoretical foundation for the task of grouping like documents and template identification and extraction.

2. Methods

2.1. Overview

The heuristic algorithm we propose, shown in Fig. 1, is composed of two modules. In line with a pilot and proof of concept, we follow the principle of using the simplest tool that will accomplish the goal and leave improvements via additional complexity and assumptions to future research.

The first module conducts grouping to find documents with similar templates. In the terms of IF this is our information scent model, where we identify our groups of highest profitability, or those most likely to contain templates. The second module takes the groups and extracts the common template, our IF enrichment step. We do not set out at this stage of the research to extract specific or all templates present but to extract the most common templates. These would have the largest benefit for use in improving subsequent NLP tasks.

After extraction of sequences via the LCS, human review is undertaken to curate the sequences and add or remove portions to obtain a canonical template signature that can the be applied to the same or a new corpus to identify instances of the template and the portions that were inserted or modified by health care providers.

2.2. Grouping

The grouping module, shown in Fig. 2, seeks to group documents with a common template. If a group of documents were derived from the same set of templates, we infer the structure, length and token count would be expected to be similar. We operate as if the converse were true. That is, if the word counts are similar, then the documents were derived from the same templates. While we recognize that this is not strictly true, we use it to form the scents of our IF scent following task. We test this assumption by measuring the percentage of groups found with more than 3 documents that have a common template as determined by human review.

We tokenize on whitespace. Punctuation is either part of the template or part of the text inserted by the clinical, and so there is no justification for separate handling of punctuation. Additionally, we make no attempt for semantic representation variations. The assumption is differing representations of key concepts would either (1) be part of the template, in which case it would distinguish a separate template, or (2) be part of the information added by a provider, where it would not affect template identification and would be handled in subsequent information extraction steps, and are not in the scope of this tool.

A further simplification we use is for making the problem computationally feasible, an IF enhancement step to satisfy our computational constraints. In place of using raw token counts, which have over 70,000 possible tokens, we use a fixed hash that reduces the dimension of the word counts to 512 bins. We use the cyclical redundancy check algorithm with a 9-bit hash. The hash key is chosen to maximize the variability in the resulting bins. It follows that if the token hashes are similar, then the many to one map reduction retains the similarity. We use hierarchical clustering,