Automated Identification of Relevant New Information in Clinical Narrative

Rui Zhang
Institute for Health Informatics
University of Minnesota
Minneapolis, MN 55455
ghan386@umn.edu

Serguei Pakhomov
Institute for Health Informatics
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455
pakh0002@umn.edu

Genevieve B. Melton
Institute for Health Informatics
Department of Surgery
University of Minnesota
Minneapolis, MN 55455
gmelton@umn.edu

ABSTRACT
The ability to explore and visualize clinical information is important for clinicians when reviewing and cognitively synthesizing electronic clinical documents for new patients contained in electronic health record (EHR) systems. In this study, we explore the use of language models for detecting new and potentially relevant information within an individual patient’s collection of clinical documents using an expert-based reference standard for evaluation. We achieved good accuracy with a heterogeneous system based on a modified n-gram language model with statistically-derived and classic stop word removal and lexical normalization, as well as heuristic rules. This technique also identified relevant new information not identified with the expert-derived reference standard. These methods appear promising for providing an automated means to improve the use of electronic documents by clinicians.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval - Retrieval models; I.2.7 [Artificial Intelligence]: Natural Language Processing – Language models, Text analysis.

General Terms
Algorithms, Design.

Keywords
Electronic health record, n-gram model, information retrieval, information redundancy, natural language processing.

1. INTRODUCTION
The mandated adoption of electronic health record (EHR) systems in the United States is intended to improve patient safety, care quality, and healthcare costs [1]. While EHR systems aim to have these benefits for clinical care, they result in fundamental changes in workflow and cognitive load for providers, which can have deleterious effects [2]. Increased information availability may also result in cognitive overload from the increasing volume of patient data. Thus, tools are needed to improve the retrieval of relevant new information from EHRs, while suppressing redundant or irrelevant information.

Clinical narrative is prevalent throughout clinical care, including EHR systems. Unlike structured information where certain information details can be lost, narrative text provides contextual information to help clinicians to more thoroughly understand medical conditions of patients and communicate with other healthcare providers [3]. When assessing a new or unfamiliar patient, clinicians must synthesize patient documents, which can be a time-consuming process. This process places a significant cognitive load on the physician, particularly when the patient is unfamiliar or complex. The process may also be inefficient, particularly when there are multiple long and complex clinical notes containing large amounts of redundant or irrelevant information. Techniques that can successfully assist clinicians with distilling relevant new information within these narratives could potentially improve the ability of providers to make therapeutic and diagnostic decisions by focusing cognitive effort towards salient information instead of culling through narrative to synthesize important information from these documents.

Several reports have focused upon automated summarization of patient narrative [4, 5] which may be viewed as one way of reducing the amount of work needed to process patient records; however, these techniques typically provide summarized narrative of a patient separate from the clinical document(s). They do not help clinicians focus on potential critical types of information “in situ” within the original document. There have also been some reports aimed at quantifying redundancy (non-novel information) in clinical text [6, 7] that demonstrate redundancy to be a significant issue within clinical texts. The task of identifying relevant new information from clinical narratives and presenting this information in a “lossless” fashion and the effect of presenting the text with visual cues remains a largely unexplored research area that may improve primary use as well as secondary uses (e.g., research and quality assurance) of patient documents [8-10]. Recently, we reported on the use of a modified dynamic programming alignment technique with lexical normalization and stop words removal for measuring redundancy in clinical texts and demonstrated good correlation (82%) with medical expert judgments [7].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IHI’12, January 28–30, 2012, Miami, Florida, USA.
Copyright 2012 ACM 978-1-4503-0781-9/12/01...$10.00.
clinicians, including visualization of relevant new information, is a potential effective solution. Reported methods and applications for visualization of clinical documents have not focused on either usability [11] or optimal presentation [12] of these texts. The goal of the current study was to investigate techniques to identify relevant new information and demonstrate visualization of relevant new information within clinical texts.

2. BACKGROUND

2.1 Similarity Measurement and Relevant New Information Identification

Semantic similarity [13, 14] has been studied as a basic concept in biomedical Natural Language Processing (NLP). Semantic similarity measures the degree to which two concepts are similar in meaning. Patient similarity is a type of similarity metric where the degree of similarity between patients represented as sets of concepts representing each case are compared [15, 16]. Measures of similarity (whether individual concepts or sets of concepts) can be commutative, as the subject and target item (concepts or sets of concepts) can be interchanged. Unlike similarity measures, identification of relevant, new (i.e., non-redundant) information has been largely unexplored. New information in a target item is information that is different from previously mentioned information or has not been mentioned previously in the subject item(s). In contrast, relevant information in a target item is information that is relevant to a particular task if it increases the likelihood of accomplishing the goal (in this case, a clinician understanding a given patient). Relevant new information is information in a target item not contained within subject item(s) relevant to a particular task, thus depending on the selection of target and subject items. For example, when comparing subject item(s) S and target item T, relevant new information of T is the information that was contained in T but not in S relevant to a task. The result is different when switching subject and target items. Therefore a relevant new information measurement is not commutative with T and S.

2.2 N-gram Model

Statistical language modeling has been widely used to identify patterns in non-clinical applications of text. An n-gram model is a type of probabilistic model typically used to predict words with the highest probability for all given previous words by using the chain rule of probability. The Markov assumption, commonly used for n-gram models, presumes that the sequence history only depends on words in local context and makes the n-gram model equivalent to an n-1 Markov model.

2.3 New Information in Clinical Narrative

Few studies [6, 7] have reported methods for quantifying redundant (not new) information in clinical notes. Both Wrenn et al. and our previous study used alignment methods and reported the degree of redundancy in clinical text. Zhang et al. used a small sliding window technique, in contrast to the original string alignment method that provided only a “macro” level measurement, not capable of identifying relevant new information at a “micro” level (such as at a sentence level). Although Zhang et al. presented a score-based method (high score representing redundant or similar and low score representing dissimilar or new information), redundancy was quantified for an entire document. This technique’s ability to identify relevant new information in a given document was not examined. Suermann et al. reported on a methodology to identify relevant information in an EHR system, but did not focus on relevant new narrative nor present a formal evaluation or algorithmic approach to this problem [17].

3. METHODS

3.1 System Design

Nine patient records were selected for this study. Our experimental design and system architecture are illustrated in Figure 1, with the system being developed with four patient records (“training records”) and the system tested on the remaining five records (“test records”). The workflow consisted of: 1) collecting patient documents and document metadata from the clinical document repository; 2) text preprocessing; 3) application of n-gram models and various enhancements trained on n previous documents to identify relevant new information of the (n + 1)th document for a given patient; 4) creation of an expert-derived reference standard with expert manual annotation; and 5) evaluation of automated method performance.

3.2 Data Preparation

Medical records from University of Minnesota Medical Center, Fairview Health Services were used in this study. These notes were extracted in text format from the Epic EHR system [18], which were created during a one-year period (12/2008 to 11/2009). Outpatient notes (i.e., office visits, allied health notes, telephone notes, results) were arranged chronologically. Institutional review board approval was obtained and informed consent waived for this minimal risk study.

Figure 1. Experimental design and system architecture.

3.3 Text Pre-processing

Since not all sentences in Epic clinical notes are well-formed (e.g., “review of system” may appear as the only text on a line or as part of an enumeration), we treated incomplete sentences or “statements” as sentences in this study. Each note was further separated into smaller chunks at a sentence/statement level.

3.4 Baseline Relevant New Information Identification Using Bigram Model

As a baseline metric, a single bigram language model was built based on all preceding unaugmented documents (e.g. from the 1st to nth document) to identify relevant new information within the target document (e.g. the (n + 1)th document). The bigram counts of each bigram model were used to classify relevant new information in the target document. Initially, the count threshold
value was set to zero for each bigram. For example, if \( C(w_2 | w_1) \) did not appear in any subject documents, then \( w_2 \) following \( w_1 \) in the target document was considered as a new word.

3.5 Manually Annotated Reference Standard
Nine outpatient clinical records with ten office visits per patient record were selected for this study. Four records were used for training and developing the system (about 6,200 sentences and statements) and five records (including one evaluated by both experts) were used for evaluation (about 9,700 sentences and statements). Two physicians were asked to identify new and clinically relevant information within each document (starting from the second document) based on all the preceding documents chronologically for each patient record using their clinical judgment. Each medical expert annotated five patient records with one record overlap with both. New information in documents was annotated with the General Architecture for Text Engineering (GATE) [19], which allows for the annotation of text and XML files through a graphical user interface (GUI), with a customized annotation schema.

In order to measure agreement between two clinician experts in the task of identifying new information, the overlap between annotations was measured in one of the nine outpatient clinical records manually annotated. Cohen’s Kappa statistic and percent agreement [20] were used to assess inter-rater reliability of the two physicians judgments at a sentence or statement level. If one or more words were marked as relevant new information by experts, the whole sentence or statement was considered as relevant new information for evaluating methods. Performance of automated methods compared to the reference standard was then measured by accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) at a sentence or statement level for the five test records.

3.6 New Information System Enhancements
In addition to the baseline, which relied bigram counts from unaltered raw text using the bigram language model (Baseline), several modifications were explored (Table 1).

Table 1. Modification of methods and examples.

<table>
<thead>
<tr>
<th>Method</th>
<th>ACC</th>
<th>SEN</th>
<th>SPE</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.507</td>
<td>0.957</td>
<td>0.444</td>
<td>0.193</td>
<td>0.987</td>
</tr>
<tr>
<td>CSW (bigram)</td>
<td>0.507</td>
<td>0.957</td>
<td>0.444</td>
<td>0.193</td>
<td>0.987</td>
</tr>
<tr>
<td>BSW (bigram)</td>
<td>0.649</td>
<td>0.942</td>
<td>0.608</td>
<td>0.250</td>
<td>0.987</td>
</tr>
<tr>
<td>LVG BSW (bigram)</td>
<td>0.654</td>
<td>0.961</td>
<td>0.611</td>
<td>0.255</td>
<td>0.991</td>
</tr>
<tr>
<td>HR (bigram)</td>
<td>0.829</td>
<td>0.889</td>
<td>0.820</td>
<td>0.456</td>
<td>0.982</td>
</tr>
<tr>
<td>SUB (bigram)</td>
<td>0.894</td>
<td>0.757</td>
<td>0.914</td>
<td>0.552</td>
<td>0.964</td>
</tr>
<tr>
<td>SUB (trimgram)</td>
<td>0.800</td>
<td>0.738</td>
<td>0.808</td>
<td>0.341</td>
<td>0.958</td>
</tr>
<tr>
<td>SUB (four-gram)</td>
<td>0.805</td>
<td>0.738</td>
<td>0.814</td>
<td>0.348</td>
<td>0.958</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count Threshold Value = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUB (bigram)</td>
</tr>
<tr>
<td>SUB (bigram)</td>
</tr>
</tbody>
</table>

Table 2. Comparison of methods with reference standard. ACC = Accuracy; SEN = Sensitivity; SPE = Specificity; PPV = Positive Prediction Value; NPV = Negative Prediction Value. Relevant new information defined when count ≤ to count threshold value.

<table>
<thead>
<tr>
<th>Methods (n-gram)</th>
<th>ACC</th>
<th>SEN</th>
<th>SPE</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (bigram)</td>
<td>0.471</td>
<td>0.678</td>
<td>0.442</td>
<td>0.144</td>
<td>0.909</td>
</tr>
<tr>
<td>CSW (bigram)</td>
<td>0.507</td>
<td>0.957</td>
<td>0.444</td>
<td>0.193</td>
<td>0.987</td>
</tr>
<tr>
<td>BSW (bigram)</td>
<td>0.649</td>
<td>0.942</td>
<td>0.608</td>
<td>0.250</td>
<td>0.987</td>
</tr>
<tr>
<td>LVG BSW (bigram)</td>
<td>0.654</td>
<td>0.961</td>
<td>0.611</td>
<td>0.255</td>
<td>0.991</td>
</tr>
<tr>
<td>HR (bigram)</td>
<td>0.829</td>
<td>0.889</td>
<td>0.820</td>
<td>0.456</td>
<td>0.982</td>
</tr>
<tr>
<td>SUB (bigram)</td>
<td>0.894</td>
<td>0.757</td>
<td>0.914</td>
<td>0.552</td>
<td>0.964</td>
</tr>
<tr>
<td>SUB (trimgram)</td>
<td>0.800</td>
<td>0.738</td>
<td>0.808</td>
<td>0.341</td>
<td>0.958</td>
</tr>
<tr>
<td>SUB (four-gram)</td>
<td>0.805</td>
<td>0.738</td>
<td>0.814</td>
<td>0.348</td>
<td>0.958</td>
</tr>
</tbody>
</table>

Table 3. Statistical results increasing the number of previous documents in the N-doc model. ACC = Accuracy; SEN = Sensitivity; SPE = Specificity; PPV = Positive Prediction Value; NPV = Negative Prediction Value.

<table>
<thead>
<tr>
<th># Pre Docs</th>
<th>ACC</th>
<th>SEN</th>
<th>SPE</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.834</td>
<td>0.756</td>
<td>0.845</td>
<td>0.398</td>
<td>0.962</td>
</tr>
<tr>
<td>2</td>
<td>0.852</td>
<td>0.733</td>
<td>0.868</td>
<td>0.434</td>
<td>0.959</td>
</tr>
<tr>
<td>3</td>
<td>0.861</td>
<td>0.720</td>
<td>0.880</td>
<td>0.451</td>
<td>0.958</td>
</tr>
<tr>
<td>4</td>
<td>0.860</td>
<td>0.715</td>
<td>0.880</td>
<td>0.455</td>
<td>0.957</td>
</tr>
<tr>
<td>5</td>
<td>0.871</td>
<td>0.689</td>
<td>0.896</td>
<td>0.477</td>
<td>0.955</td>
</tr>
<tr>
<td>6</td>
<td>0.866</td>
<td>0.703</td>
<td>0.890</td>
<td>0.478</td>
<td>0.954</td>
</tr>
<tr>
<td>7</td>
<td>0.862</td>
<td>0.695</td>
<td>0.887</td>
<td>0.480</td>
<td>0.951</td>
</tr>
<tr>
<td>8</td>
<td>0.885</td>
<td>0.653</td>
<td>0.918</td>
<td>0.527</td>
<td>0.950</td>
</tr>
<tr>
<td>9</td>
<td>0.883</td>
<td>0.636</td>
<td>0.920</td>
<td>0.543</td>
<td>0.944</td>
</tr>
</tbody>
</table>
4. RESULTS

4.1 N-gram Models Performance Evaluation

Cohen’s Kappa coefficient of two annotators for the overlap clinical documents was 0.65 and percent agreement was 0.94. To determine the threshold TFIDF value, we tested the first three cutoff values of 2E-6, 4E-6, 6E-6 and chose 2E-6 due to similar performance. Table 2 shows performance for different methods. The use of bi-grams with a count threshold of zero, addition of lexical normalization, removal of both stop word types, as well as heuristic rules including section-specific rules, resulted in best performance.

4.2 N-doc Model Evaluation

Table 3 shows the accuracy, sensitivity, specificity, PPV and NPV as \( N \) increases from 1 to 9. Overall accuracy and PPV increased with increasing document numbers, resulting in decreasing sensitivity with increasing documents in the model.

4.3 Relevant New Information Visualization

Figure 2 shows example screen shots of clinical notes highlighted by these methods in comparison to the expert reference standard. Relevant new information at a word level is highlighted as green in comparison to reference standard relevant new information in purple. In Sec1, formatting and signature were not marked for both (True Negative (TN)), and the first paragraph was marked in both as relevant new information (True Positive (TP)). The automated method wrongly marked the second paragraph, which is a False Positive (FP). In Sec2, relevant new information about MUSCULOSKELETAL was marked in both (TP). But another piece of relevant new information “Negative for temperature intolerance, skin/hair changes” was marked by the automated method (FP). In Sec3, the diagnosis was correctly marked, however, the plan was not marked as relevant new information by automated method (FN).

5. DISCUSSION

This study focuses on the unexplored and increasingly important topic in clinical informatics: identification and visualization of relevant new information in medical texts. In this study, we explore techniques to detect relevant new information in clinical notes. We developed an expert-based reference standard and used it to evaluate several approaches, including baseline n-gram techniques, as well as several enhancements such as rule-based and statistical knowledge-free approaches. Our study shows that the content words, knowledge-base rules (e.g. formatting and noise removal, and section rules in the clinical notes) are important features that need to be included when distinguishing between relevant new versus redundant information.

We observed that heuristic rules helped to improve system performance, including section-specific rules. Informal analysis of the removed content shows that noise is introduced by structural attributes of clinical notes. For example, most clinical reports in our system include visit information located at the head of the document, such as visit date, time, encounter number, provider, previous visit et al as well as the signatures at the end of some subsections to indicate the document have been reviewed or recorded. Prior to removing these items, our n-gram model marked them as relevant new information since some parts of these items typically change from visit to visit. Other examples include the details of medications, such as medication class, route and special instructions. While this medication information may change slightly over time, it was judged by physician annotators as irrelevant. Section headings constitute another cue indicative of redundant or irrelevant information. For example, the content of
the “Past Medications” and “Allergies” sections was always marked by annotators as redundant information; in contrast, the “Follow-up” section was always marked as relevant new information. A particularly interesting example consists of two documents showing exactly the same sentence, “patient will keep a food record and return in 2 weeks”, marked by experts as relevant new information.

We also observed that the two medical experts sometimes showed slightly different views on annotating relevant new information. Both annotators failed to mark some of the new information. We also found that one expert annotated more carefully than the other resulting in fewer missing values. For example, one annotator failed to mark “body mass index is ...” as relevant new information in contrast to our automated approach. Another example is the statement “Height: 5’10’’” stated multiple times in previous notes that was marked as relevant new in the reference but redundant by our methods, accounted as FN. So it is necessary to refine the reference standard for further study by adjudicating both annotations by a third expert or by refining instructions to annotators in subsequent experiments.

Our methods are currently not designed to identify relevant new information at the semantic level. Due to this limitation, our approach is sensitive to such variation as the use of acronyms and word order changes. Another interesting observation is that trigram and four-gram approaches performed worse than the bigram model. This is not surprising because the models were trained on a relatively small corpus (e.g. about 2,000 sentences and statements). Thus these data may have been too sparse for higher than bigram order modeling.

To investigate the effect of the number of previous documents used for modeling, we varied the number of preceding documents from 1 to 9. The PPV increased on average by 15% with the number of preceding documents. Another interesting observation is that higher order changes. Another interesting observation is that trigram and four-gram approaches performed worse than the bigram model. This is not surprising because the models were trained on a relatively small corpus (e.g. about 2,000 sentences and statements). Thus these data may have been too sparse for higher than bigram order modeling.

To investigate the effect of the number of previous documents used for modeling, we varied the number of preceding documents from 1 to 9. The PPV increased on average by 15% with the number of preceding documents. Another interesting observation is that trigram and four-gram approaches performed worse than the bigram model. This is not surprising because the models were trained on a relatively small corpus (e.g. about 2,000 sentences and statements). Thus these data may have been too sparse for higher than bigram order modeling.

Overall, this study of automated visualization of relevant new information via highlighting showed that this was a simple and effective way to present information to physicians when they review complex medical documents. The results of our exploratory study are significant in the context of the development of next generation EHRs that should take into account human factors such as information overload in text, which can affect patient safety and quality of care.

6. ACKNOWLEDGMENTS

This research was supported by the American Surgical Association Foundation Fellowship (GM), UMN Institute for Health Informatics Seed Grant (GM & SP) and National Library of Medicine (#R01LM009623-01) (SP). The authors thank Fairview Health Services for ongoing support of this research.

7. REFERENCES