Abstract
An accurate problem list plays the key role of a problem-oriented medical record, which plays a significant role in improving patient care. However, the multi-author, multi-purpose nature of problem list makes it a challenge to maintain, and a single list is difficult, if not impossible, to satisfy all the needs of different practitioners. In this paper, we propose using machine generated problem list to assist a medical practitioner to review a patient’s chart. The proposed system scans both structured and unstructured data in a patient’s electronic medical record (EMR) and generates a ranked, recall-oriented problem list grouped by body systems. Details of each problem are readily available for the user to assess the correctness and relevance of the problem. The user can then provide feedback to the system on the trustworthiness of each evidence passage retrieved, as well as the validity of the problem as a whole. The user-specific feedback provides new information the system needs to perform active learning to learn the user’s preference and produce personalized, and/or domain-specific problem lists.

Introduction
Since the publication of Dr. Lawrence Weed’s seminal paper on problem-oriented medical record (POMR) in 1968 (Weed, 1968), organizing medical records around the problem list has become a commonly accepted practice. The meaningful use program launched by the federal government also requires the use of problem list in electronic health records (EHR) systems. However, creating and maintaining an accurate problem list in the traditional way has proven to be challenging (Campbell, 1998). To address this, a new manual documenting model has been proposed by physicians (Mehta, 2014), and automated problem list generation methods have emerged from the natural language processing (NLP) community (Meystre & Haug, 2005, 2008), (Solti. et. al. 2008), (Danielle, 2014), (Devarakonda & Tsou, 2015). While the automated methods can extract reasonably accurate disorders and findings from EMRs, we feel two fundamental issues of generating useful problem list can be better solved by building a cognitive assistance system. First, what should be on the problem list? While there are a variety of authoritative sources offering problem list definitions (AHIMA, 2008), there is no single definition of problem list that applies across all specialties. An internist may appreciate a comprehensive problem list, whereas a nephrologist may find that the same list contains too many irrelevant findings but not enough details pertaining to his specialty. Second, a single problem list does not take personal preference into account. Some practitioners prefer a succinct list containing only the active problems, while other practitioners find significant disorders resolved in the past to also provide valuable information. In this paper, we propose a cognitive assistant system that,

- generates recall-oriented problem lists automatically
- merges clinically similar problems
- groups problems based on standard taxonomy
- ranks problems based on strength of the evidences found in the EMR
- provides evidences of each problem for human judgement
- collects user inputs and preferences, and improves subsequent problem lists using active learning

Automated Problem List Generation
Using NLP and machine learning to generate problem lists automatically from EMRs has started to draw the attention of the research community in recent years. For example, University of Utah (Meystre & Haug, 2005) has achieved a recall of 90% and precision of 69% from extracting a predefined set of 80 selected problems; University of Washington (Solti et. al. 2008) generalized the approach to analyze...
one week’s worth of cardiology ambulatory progress notes and extract problems found in certain sections within those notes, achieving a recall of 88% and precision of 66%. To the best of our knowledge, the IBM Watson-based system (Devarakonda & Tsou, 2015) is the first successful automated problem list generation system that extracts an open-ended list of patient’s medical problems from the entire longitudinal EMR, which often contains hundreds of notes in the course of several years. The current IBM Watson-based system has a recall of 84% and precision of 53%. As there is no shared corpus and the accuracy evaluation methodology varies, it is challenging to compare the accuracy among the systems at this point.

Ground-Truth Generation
358 longitudinal EMRs are currently used to train (278 EMRs) and test (the remaining 80) our system. Each EMR was independently reviewed by two fourth year medical students, resulting in two separate problem lists that were then adjudicated together with an MD. As a longitudinal EMR is a rich source of information, reviewing a single EMR and generating a comprehensive problem list requires 30 minutes in average per annotator – apparently a pure supervised learning method will not scale up.

Inter-annotator Agreement
Using the adjudicated result as ground-truth, the inter-annotator agreement accuracy are shown in Table 1.

<table>
<thead>
<tr>
<th>Annotator</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>334</td>
<td>27</td>
<td>98</td>
<td>0.78</td>
<td>0.93</td>
<td>0.84</td>
</tr>
<tr>
<td>B</td>
<td>461</td>
<td>53</td>
<td>61</td>
<td>0.88</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>C</td>
<td>427</td>
<td>31</td>
<td>44</td>
<td>0.91</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>Sum/Avg</td>
<td>1122</td>
<td>111</td>
<td>203</td>
<td>0.86</td>
<td>0.92</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 1: Inter-Annotator Agreement

Note that the recall is comparable to our system but the precision is significantly higher.

Error Analysis and Motivation
As the system recall is similar to human experts, we focused our error analysis on finding the sources of false-positives. The major sources of false-positives are shown in Table 2.

<table>
<thead>
<tr>
<th>Error Category</th>
<th>Error Type</th>
</tr>
</thead>
</table>
| Subjective     | S1. transient problem  
S2. redundant problem  
S3. hypothetical problem |
| Objective      | O1. negation error  
O2. abbreviation expansion error  
O3. noise (template / checkbox) |
| Not an error   | N1. ground-truth error |

Table 2: Types of False-Positive Problems

The first category contains problems that are transient, minor, resolved in the past, or already covered by another problem, and problems that were considered or even diagnosed by a physician but the supporting evidence found in the EMR was too weak to be certain. The current system does not model this well enough, but at the same time this category is well suited for a cognitive system as human judgment plays an important role. The second category contains actual errors the system is making. Clinical notes in EMR are nosier than most text, and we have made significant progress to reduce errors in this category, but this remains work in progress. The third category is interesting. Although our ground-truth problem lists have been carefully reviewed and vetted, our system still finds real problems that are overlooked by the human experts. After using the system generated problem list to help vet the ground-truth, approximately one new problem has been added to every two EMRs (an average EMR in our training data has 6 problems). In the clinical setting, a time-strapped physician will have less time than our annotators to review the charts and will benefit from a comprehensive automated problem list. All false-positive problems from one of our testing EMRs are shown in Table 3.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Error Type</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal pain</td>
<td>S1</td>
<td>finding</td>
</tr>
<tr>
<td>Anxiety state</td>
<td>S1</td>
<td>diagnosed, controlled</td>
</tr>
<tr>
<td>Headache</td>
<td>S1</td>
<td>finding</td>
</tr>
<tr>
<td>Anorectal fistula</td>
<td>S1</td>
<td>resolved (surgeries)</td>
</tr>
<tr>
<td>Epigastric pain</td>
<td>S1</td>
<td>finding (related to abdominal pain)</td>
</tr>
<tr>
<td>Infectious mononucleosis</td>
<td>S1</td>
<td>acute / resolved</td>
</tr>
<tr>
<td>Anal fissure</td>
<td>S2</td>
<td>related to fistula</td>
</tr>
<tr>
<td>Crohn's disease</td>
<td>S3</td>
<td>questionable</td>
</tr>
<tr>
<td>Asthma</td>
<td>O3</td>
<td>&quot;Do you have a history of asthma?&quot;</td>
</tr>
<tr>
<td>Chronic sinusitis</td>
<td>N1</td>
<td>true problem</td>
</tr>
<tr>
<td>Allergic rhinitis</td>
<td>N1</td>
<td>true problem</td>
</tr>
</tbody>
</table>

Table 3: False-Positive Problems from an EMR

In the original ground-truth this patient has 9 problems. Auto-generated problem list identified all of them (recall = 1.0) and 11 more (precision = 0.45). If we exam the 11 false-positives closely, many of them provide useful information and may be considered as true problems by some practitioners.

Cognitive Assistant Interface
The high percentage of subjective judgements observed in our error analyses gives us strong motivation to build the system as a cognitive assistant. The system first presents a
recall-oriented problem list, trained using supervised learning, to the user. The list is ranked and clustered to reduce the cognitive load of the user. Evidence of each problem is presented when the problem is selected. After reviewing the evidence passages, the user can provide feedback to the system, and the feedback will be used to train the system to generate subsequent problem lists based on the user’s preference.

**Problem List User Interface**

Screenshots of the prototype system are shown below. Figure 1 shows the initial problem list ordered by the system generated confidence score. Higher score indicates stronger evidence can be found in the EMR, and therefore the problem is more likely to be a true and relevant problem of this patient.

![Figure 1: Problem List](image1)

Figure 2 shows the same list now grouped by a SNOMED CT category (Disorder of body system). When related problems are clustered, redundant findings become less bothersome and in some cases actually provide additional information.

![Figure 2: Clustered Problem List](image2)

Figure 3 shows how evidence passages are displayed when a problem is selected. A problem is often discussed in multiple notes. In addition to the metadata, such as the note date and provider name, the system also infers the note type and note section to provide the user more context without opening-up the actual note. Currently only evidence found in clinical notes are displayed. Other types of evidence, for example, relations to active medications and abnormal test results, are used by our model to score the problems but not displayed in the current UI. While reviewing evidence for a specific problem, the user can give feedback for each evidence passage retrieved by the system by clicking on the checkboxes shown in Figure 3 to indicate if the evidence retrieved supports the problem indicated. After reviewing all evidence for a selected problem, the user can then provide feedback on the problem as a whole by clicking on the checkboxes shown in Figure 1 to indicate if a problem is considered incorrect, or correct but irrelevant to the particular provider.

![Figure 3: Evidences of Selected Problem](image3)

**Active Learning**

The initial problem list is generated using supervised learning. The drawback of using supervised learning is, 1.) Ground-truth is time consuming to generate 2.) Ground-truth represents one particular view / definition of the problem list, which is not necessarily aligned with every practitioner. Both problems can be overcome by active learning, and the proposed cognitive assistant UI provides a natural way to interactively query the user to obtain the desired outputs at new data points. As the problems are ranked by the system’s confidence, data collected on this actively selected set is most informative, in terms of error-based active learning where the goal is to reduce the predictive error.

Problems marked by a user as incorrect or irrelevant will be hidden from view the next time the same user opens the same EMR — no learning required. This user-specific data will also be used to train a user-specific model that learns the user’s preference and ranks problems in other EMRs accordingly. Problems marked by multiple users as incorrect or irrelevant will be used as ground-truth to improve the initial problem lists generated using supervised algorithms.

**Conclusion and Future Work**

A patient is often treated by multiple providers, and this makes problem lists inherently multidimensional. A single, static problem list, whether manually maintained or automatically generated, will not satisfy all providers’ needs.
Building on top of our state-of-the-art problem list generation algorithm, we propose a cognitive assistant that uses active learning to build specialty-specific and personalized problem lists. The proposed system is summarized in Figure 4. The design goal of the initial problem list is to help physicians to not miss any problems, i.e., a recall-oriented system. This is because it is much harder for a physician to create a comprehensive problem list than to verify if a particular problem is correct and relevant. To reduce the cognitive load on the user, the list can be easily sorted by any attribute and grouped based on standard taxonomy, and evidence of the problems are readily available for the user to review. The cognitive assistant not only presents the initial lists, but also actively requests user inputs for problems with higher ranking to obtain the most informative training data, and then employs the feedback to produce more accurate, and more personalized problem lists.

The core problem list generation engine itself is a work in progress, and any improvement made in the core algorithm will have direct impact on the cognitive assistant. Deploying the proposed cognitive assistant to the annotators and practitioners will also help gather instructive training data that is costly to obtain otherwise. Currently only evidence found in clinical notes are presented in the user interface when a problem is selected. However, physicians often consider test results to be the primary data when other evidence from clinical notes are controversial. Many types of structured data, including lab tests, are considered as a whole in the problem list model, but the instance level accuracy is not yet accurate enough to be displayed as evidence. Higher accuracy in relation detection is an active area we are pursuing. The user interface, as of the time of writing, is a proof-of-concept prototype. To support better personalized features, a full-fledged user management system is needed.

The concept of applying active learning in recommender systems has shown great success in many large scale consumer facing systems, such as Netflix, Amazon, and Facebook. Similar to problem lists, recommender systems often have an ill-defined or open-ended objective - to find what it is the user likes to see. With the proposed cognitive extension to our problem list generation engine, the system provides straightforward mechanisms for providers to review patients’ charts, keeps the users in the feedback loop, and automatically improves the problem lists’ accuracy based on the individual and aggregated response.

References


