tapraid2/ami-jamia/ami-jamia/ami00607/ami2042d07z | jadhavr | S=1 | 10/24/07 | Art: 458 | ce: 31

Research Paper ■

The Evaluation of a Temporal Reasoning System in Processing Clinical Discharge Summaries

AQ: 1

LI ZHOU, PHD, BMED, SIMON PARSONS, PHD, GEORGE HRIPCSAK, MD, MS

Abstract **Context:** TimeText is a temporal reasoning system designed to represent, extract, and reason about temporal information in clinical text.

Objective: To measure the accuracy of the TimeText for processing clinical discharge summaries.

Design: Six physicians with biomedical informatics training served as domain experts. Twenty discharge summaries were randomly selected for the evaluation. For each of the first 14 reports, 5 to 8 clinically important medical events were chosen. The temporal reasoning system generated temporal relations about the endpoints (start or finish) of pairs of medical events. Two experts (subjects) manually generated temporal relations for these medical events. The system and expert-generated results were assessed by four other experts (raters). All of the twenty discharge summaries were used to assess the system's accuracy in answering time-oriented clinical questions. For each report, five to ten clinically plausible temporal questions about events were generated. Two experts generated answers to the questions to serve as the gold standard. We wrote queries to retrieve answers from system's output.

Measurements: Correctness of generated temporal relations, recall of clinically important relations, and accuracy in answering temporal questions.

Results: The raters determined that 96.9% of subjects' 295 generated temporal relations were correct and that 96.5% of the system's 995 generated temporal relations were correct. The system captured 79.2% of 307 temporal relations determined to be clinically important by the subjects and raters. The system answered 83.7% of the temporal questions correctly.

Conclusion: The system encoded the majority of information identified by experts, and was able to answer simple temporal questions.

■ **J Am Med Inform Assoc.** 2007;xx:xxx. DOI 10.1197/jamia.M2467.

Introduction

Temporal information is an essential component of medical records.[1–3] Effective use of temporal information can help health care providers and researchers study and understand medical phenomena such as the progress of a disease, the patient's clinical course, and the clinician's reasoning. Many medical information systems use temporal information to answer time-oriented clinical queries.[4,5] to predict future consequences based on the current status of a patient,[6] to explain the possible causes of a given clinical situation,[7] and to recognize temporal patterns and create an abstract view of the data.[8–10] However, most previous studies have focused on temporal information stored in structured clinical databases.

Medical text, such as progress notes, discharge summaries and radiology reports, contain important clinical findings[11,12] (e.g., evolution of a disease and its corresponding treatment at the different stages). Medical natural language processing (NLP) systems[11] have been developed for the extracting, structuring and encoding clinical information from the text. Automatically discovering temporal relations among medical events stated in the text will dynamically link the extracted clinical information, which in turn will facilitate subsequent processing, such as conducting information retrieval and text summarization, inferring other relations (e.g., causal and explanatory relations), and detecting clinical practice patterns. In addition, having time attached to medical events will make extracted clinical information much more understandable to users. Despite the recent developments in biomedical NLP, temporal information in medical text has not been widely exploited for the support of temporal reasoning tasks.[1]

Affiliations of authors: Department of Biomedical Informatics (LZ, GH), Columbia University, New York, NY; Clinical Informatics Research and Development (LZ), Partners HealthCare, Boston, MA; Department of Computer and Information Science (SP), Brooklyn College, Brooklyn, NY.

This work was funded by National Library of Medicine (NLM) "Discovering and applying knowledge in clinical databases" (R01 LM006910).

The authors thank Carol Friedman for the use of MedLEE (NLM support R01 LM007659 and R01 LM008635). The authors also thank John Chelico, Amy Chused, Peter Hung, Xin Liu, Daniel Stein, and Ying Tao for conducting the system evaluation.

Correspondence and reprints Li Zhou, PhD, BMed, Clinical Informatics Research and Development, Partners HealthCare, 93 Worcester Street, 2nd Floor, Wellesley, MA 02481; e-mail: <lzhou2@partners.org>.

Received for review: 04/03/07; accepted for publication: 09/20/07.

A few studies[13,14] presented methods on modeling and processing temporal information in medical narrative reports. They applied natural language processing and medical knowledge to obtain a representation of time for the narrated medical events and to order these events chronologically. However, these systems' performance for such tasks was not clear. Recent research[15,16] in this area embraces probabilistic and machine learning approaches.

In order to process temporal information in clinical narrative data, researchers in biomedical informatics face many challenges.[1] Evaluating temporal NLP systems is critical to progress. In this paper, we present our evaluation of a comprehensive temporal reasoning system called TimeText in processing discharge summaries. In the background section, we will introduce the TimeText system and briefly describe our previous evaluation of the components of the system. This study is an overall evaluation of the entire system. We assess the system's performance on ordering medical events and answering queries of interest, using experts as judges. We discuss its strengths and weakness as well as providing insights in building such systems.

Background

The TimeText System

We developed a systematic temporal reasoning methodology and a corresponding system, called TimeText, for handling temporal information in electronic clinical reports, with the aim of improving biomedical information applications such as information retrieval, medical errors detection, and syndromic surveillance. TimeText is an end-to-end system that mainly consists of four components.[17] Figure 1 shows an overview of the system. It formalizes temporal assertions stated in clinical discharge summaries in the form of a Temporal Constraint Structure (TCS).[18] A temporal information recognition and normalization program, named TCS tagger, was developed to implements the TCS. TimeText uses the MedLEE[19,20] natural language processor to parse the non-temporal information (i.e., medical events). MedLEE is a comprehensive NLP system developed at Columbia University Medical Center that reads textual clinical reports and generates structured information. TimeText also includes a knowledge-based subsystem[21] which uses medical and linguistic knowledge for handling implicit temporal information and resolving issues such as granularity and uncertainty. After extracting and structuring temporal information and medical events, a computational mechanism called a Simple Temporal Constraint Satisfaction Problem (STP) was adopted for further reasoning about temporal relationships in clinical reports.[22] TimeText models temporal assertions about medical events in a discharge summary as an STP and produces the derived temporal information. The system-generated information can be used to answer questions about the time of events and the temporal relation between pairs of events. Examples included, "When was the operation conducted?" and "Did the infection occur before or after this operation?" The TimeText system architecture and detailed description of each component have been published.[17]

The TimeText system mainly consists of four components, including 1) a Temporal Constraint Structure (TCS)[18] for representing various temporal expressions and the TCS tagger; 2) an integration component with an existing medical NLP system (MedLEE)[19,20] for processing clinical information; 3) a knowledge-based subsystem[21] which uses medical and linguistic knowledge for handling implicit and uncertain temporal information; and 4) a formal temporal model[22] based on simple temporal constraint satisfaction problem for reasoning about related information in clinical reports.

Review of Previous Formative Evaluations of the TimeText Components

We conducted evaluations testing the suitability and feasibility of models and methodologies for the major components of TimeText while the system was in development. Evaluation of the Temporal Constraint Structure (TCS)[18] showed that 1961 out of 2022 (97%) temporal expressions identified in 100 discharge summaries were effectively modeled using the TCS. Note that medical dosing and some temporal adjectives and adverbs (e.g., "occasional" and "chronic") were not counted. The natural language processor MedLEE[19,20] has been used by investigators at Columbia University Medical Center since 1995. It has been applied to most types of medical text, including radiology reports, discharge summaries, pathology reports and visit notes, and achieved great accuracy across this wide range of medical text.[19,23,24] We have tested and demonstrated that most of the temporal assertions found in electronic discharge summaries can be modeled as a simple temporal constraint satisfaction problem (STP),[22] including a description of fifteen special issues on encoding and how we dealt with them.

In our previous work, we addressed fundamental issues encountered at different linguistic layers and modeling processes, conducted system architecture design, and carried out some formative evaluations which shaped the course of subsequent integration of the components. In this paper, we evaluate the overall functionality and performance of the system after all the components were put

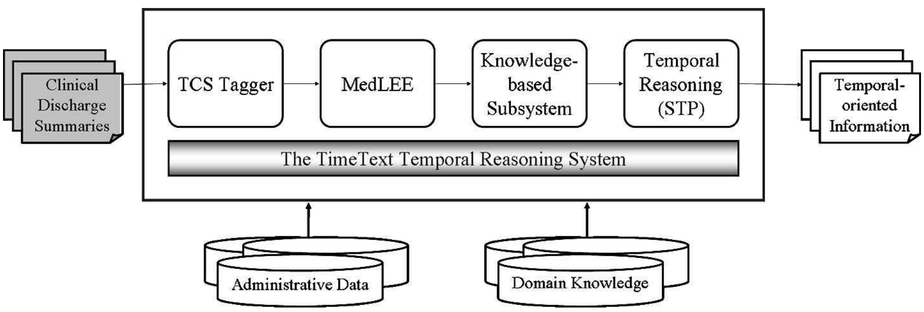

Figure 1. •••

together and a comprehensive temporal reasoning system for clinical reports was developed. In particular, we assess the accuracy of the system on ordering medical events and on answering temporal questions. We also discuss critical issues encountered during the evaluation.

Methods

The evaluation of the TimeText temporal reasoning system in processing clinical discharge summaries consists of two parts: a verification of its output temporal constraints and an assessment of its performance in answering clinical queries. We randomly selected 20 discharge summaries from a clinical data repository at Columbia University Medical Center, which contains 300,000 reports from 1989. Six physicians who have biomedical informatics training served as evaluation domain experts and helped with the evaluation. Four of them are biomedical informatics postdoctoral fellows and another two are biomedical informatics PhD candidates. None of them participated in the design or development of the TimeText system.

Part I: Verification of Output

Due to time limitations, only the first fourteen discharge summaries were used to assess the accuracy and coverage of the system-generated temporal relations between pairs of medical events (see Figure 2; Note that readers may also refer to Figure 4, which presents a summative illustration for both evaluation methods and results). From each discharge summary, five to eight clinically significant events were selected by one author (LZ, a biomedical informatics PhD candidate with a medical degree), based on the following criteria: the events included 1) reference events (e.g., admission and discharge) for the purposes of assessing the system's capability of detecting situations such as whether an event occurred before, during, or after hospitalization, because this function might be helpful for detecting medical errors; and 2) encounter-based patient-specific medical events for the purposes of assessing whether the system can capture these events as well as related temporal references and whether the system can infer correct temporal relationships. The latter included different types of medical events such as the patient's chief complains and symptoms (e.g., chest pain), important examinations and procedures (e.g., cholecystectomy), major medications (e.g., Lasix), and leading diagnoses (e.g., esophageal cancer), which were largely critical to the patient's hospital encounter. In total, 92 medical events were used for evaluation. Appendix 1, available as a JAMIA online-only data supplement at www.jamia.org, shows a simple example in the questionnaire, including a discharge summary, selected medical events, the orderings of these events generated by the system and physicians, querying questions, and the corresponding answers, which will be described in Part II. Appendix 2, available as a JAMIA online-only data supplement at www.jamia.org, shows all of the 92 selected medical events.

We model the time over which an event occurs as an interval.[22] Each interval has a start point and a finish time point and the start is never after the finish. The TimeText temporal reasoning system generated temporal relations between endpoints of paired medical events. All of the six physicians participated in this part. We asked two physicians (one is a postdoctoral fellow who completed an internship in Internal Medicine and another is a PhD student who was an astronaut physician) to serve as subjects to manually generate temporal relations for endpoints of these medical events; one encoded nine reports and another encoded five reports. Before the manual encoding, training was provided to the two subjects, including encoding instructions and a concrete example. The subjects did not attempt to exhaustively list all the temporal relations about each medical event, which would have been prohibitively time-consuming, but instead listed clinically important ones in regard to each specific patient case.

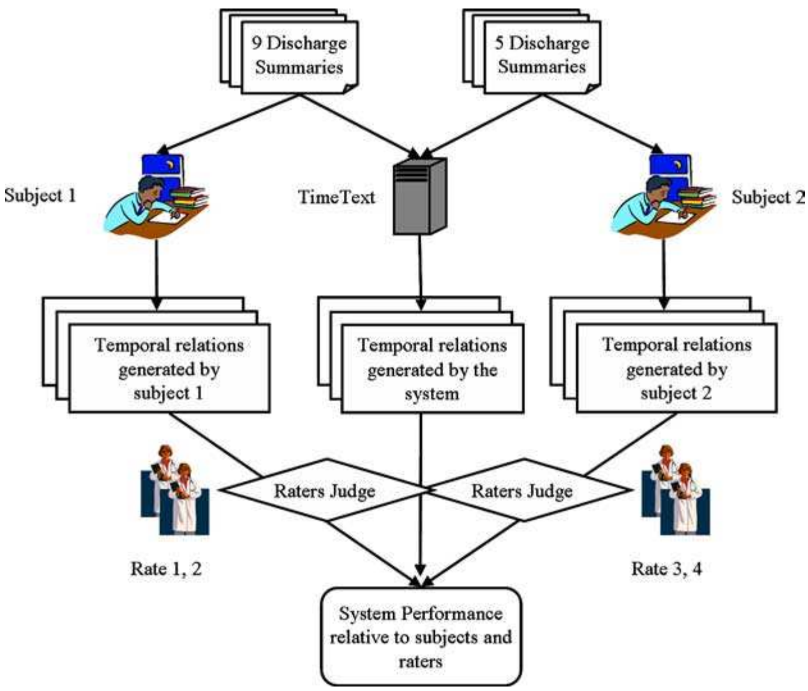

Figure 2. •••

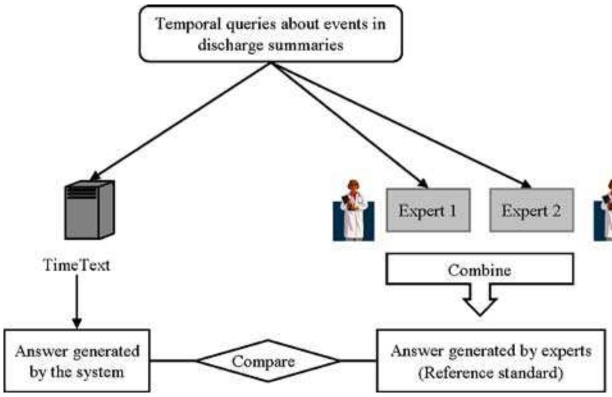

Figure 3. •••

In order to compare the performance on ordering medical events between the system and the subjects, both the system and subject-generated results were presented, blindly, to four other physicians (raters). A pair of raters reviewed the results generated by one subject and the system. They assessed the accuracy of these relations. They further identified other clinically important temporal relations that the subjects missed. Based on subject-generated results, after incorrect relations were removed and missing relations were added, a new set of relations were then generated. This new set served as a reference to assess the system's ability to identify clinically important temporal relations. Because inferring complex temporal relations was difficult even for our domain experts (subjects and raters), disagreement between the system and the experts was studied in more detail by the investigators to ascertain which was correct.

We calculated the correctness of generated temporal relations, as well as recall of the system for generating clinically important relations. We further studied spurious temporal relations (relations that were not really there) and misinterpreted temporal relations. We analyzed the sources of disagreement between the system and the subjects.

Part II: Performance in Answering Time-oriented Clinical Questions

We assessed the ability of TimeText to answer time-oriented
F3-4 clinical questions (Figure 3 and Figure 4). All twenty discharge summaries were used in this part. For each report, one author (LZ) created five to ten clinically plausible temporal queries about medical events in the reports. Similar to evaluation Part I, these queries related to the patient's predominant clinical findings. In particular, the queries might ask when an event occurred (absolute date/time); how long did an event last (duration); or whether an event occurred during hospitalization. Appendix 3, available as a JAMIA online-only data supplement at www.jamia.org, lists all the time-oriented querying questions for evaluation Part II. Two physicians, who also were subjects in Part I, served as experts to generate answers to the queries. For disagreement, we asked the experts to modify responses on the basis of the others' opinions. The modified responses were collated and returned to the experts for further modification. The process was repeated until a consensus was achieved or there were no further changes. The responses that were agreed upon then served as the reference standard. The authors wrote simple queries to retrieve answers from the system-generated temporal relations of medical events. They compared the answers generated by the system to the reference standard.

To assess the system performance, we calculated the accuracy (the proportion of correct responses) and ascertained the causes of the errors. We also calculated inter-rater disagreement to assess our experts' reliability on temporal queries.

Results

Part I: Verification of Output

Physician Performance and Reference Standard

Table 1 and Table 2 show the performance of the subjects in T1-2
generating temporal relations between endpoints of pairs of medical events. Figure 4 illustrates the results graphically. Two physicians (subjects) encoded 295 temporal relations about the 92 selected clinically important events. Four other physicians (raters) examined these relations, found 4 spurious relations, corrected 5 misinterpreted relations, and added 16 missing temporal relations that they considered clinically significant. In summary, 307 (295−4−5+5+16) clinically important temporal relations about 92 medical events were identified and they served as a reference stan-

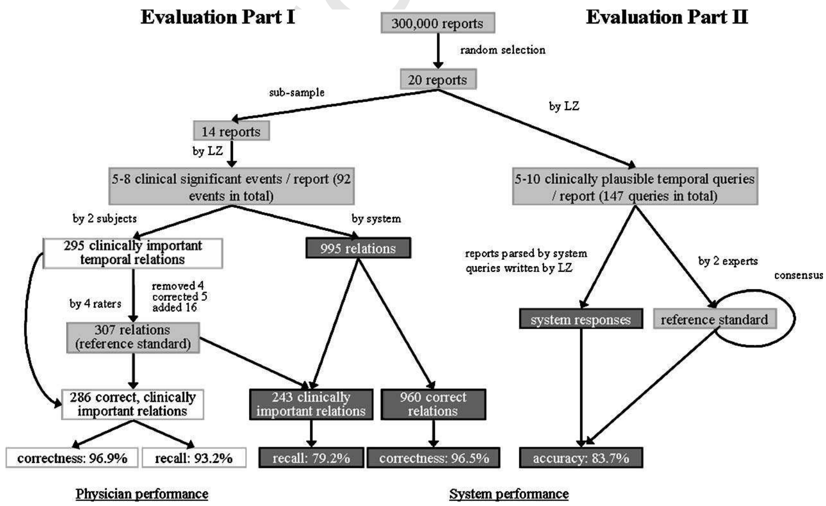

Figure 4. •••

Table 1 ■ Temporal Relations Generated by the Subjects versus the System

	Subjects	System
Total generated relations	295	995
Correct relations	286	960
Incorrect relations (inferred incorrectly)	5	30
Spurious relations (no evidence in report)	4	5
Correct relations in common with the reference standard of clinically important relations	286	243

dard to assess the system's recall. Of the 614 endpoints referenced in these relations (two per relation), 84.7% were start points of medical events and 15.3% were finish points.

Raters determined that 96.9% (286 out of 295; 95% CI: 94.3–98.4) of subjects' relations were correct (Table 2). The subjects captured 93.2% (286 of 307; 95% CI: 89.8–95.5) of the clinically important temporal relations, but because subjects helped to determine the reference standard, this result is likely an overestimate.

Error Analysis on Physician Performance

We analyzed the incorrect relations generated by subjects. There were several types. Some errors were obvious. For example, one patient was admitted for sickle cell crisis. The finish of the event should be after admission, but the annotator wrote "before." In another case, it was stated in the report that "he underwent a V-Q scan on 8/23" and that the admission was on 8/24, so that V-Q scan occurred before admission. However, the subject encoded that the V-Q scan occurred after admission. In another case, "The patient cleared of nausea and vomiting" was after using "Thorazine," while the subject encoded it the other way around.

The subjects also made spurious temporal assertions. For example, based on the statement "he experienced pancreatitis secondary to the IV Pentamidine," the subject inferred that "the finish of the IV Pentamidine was after the finish of pancreatitis." There was no evidence in the report to support this assertion.

The subjects also missed 16 temporal relations which the evaluators considered important. For example, in a report, the patient had a resection of petrous apex meningioma. His postoperative course was complicated by hemiparesis. The temporal relation between the operation (resection of petrous apex meningioma) and its complication (hemiparesis) was missed.

System Performance

Table 1, Table 2, and Figure 4 show the performance of the system in generating temporal relations between medical events. The system generated 995 temporal relations about these 92 medical events. The raters determined that 5 relations were spurious and 30 were incorrect, so that 96.5% (960 out of 995; 95% CI: 95.2–97.5) were correct. Compared to the reference standard of clinically important relations, the system missed 64 temporal relations and achieved a recall of 79.2% (243 of 307; 95% CI: 74.3–83.3). The system captured 85.8% of start points but only 42.6% of finish points that were in the reference standard of clinically important relations.

Error Analysis on System Performance

We examined the missed temporal assertions. The majority were due to finish points of medical events that were not constrained. The major reason for the errors was misplaced contents in the original reports. For example, physicians sometimes wrote the patient's current problems or current treatments in the "history of present illness" section. In one report, there was no hospital course section at all and medical events occurring during hospitalization were stated in the "history of the present illness" section.

Performance Comparison of the Physicians and the System

Of the five incorrect relations that were generated by subjects, the system generated three correctly. For example, in a report, Cefuroxime was given after the patient developed papular rash. The system successfully ordered these two events. However, the subject encoded that the start of rash was after Cefuroxime. In addition, of the 21 relations that were missed by subjects, the system captured eight.

Part II: Performance in Answering Time-oriented Clinical Questions

Inter-rater Agreement and Reference Standard

Overall, in 20 discharge summaries, 147 temporal questions about medical events were generated. Eighteen questions related to specific dates or times (for example, when did this patient have a skin graft?). Eight questions related to durations (for example, how long did diarrhea last?). Others were yes/no questions (did pancreatitis occur after pentamidine; did the patient vomit before using Thorazine; did the patient stop vomiting after using Thorazine?). The experts disagreed on 17 answers (raw inter-rater agreement: 88.4%). Four of these questions were related to durations and others were yes/no questions. A reference standard was established after the experts achieved an agreement upon their responses.

System Performance on Answering Temporal Queries

The answers generated by the system were compared to the reference standard. For yes/no and dates/times questions, an exact match was required. For questions related to durations, range estimation was allowed. For example, the answers were considered to match if the physician's answer was "3 days" while the system estimated "2–4 days." However, the system's answer was considered incorrect if the range did not cover the exact duration. In addition, if the

Table 2 ■ Performance Comparison of the Subjects and the System

	Subjects		System	
Metric	Derivation	Value (95% CI)	Derivation	Value (95% CI)
Correctness of relations	286/295	0.969 (0.943–0.984)	960/995	0.965 (0.952–0.975)
Recall of clinically important relations	286/307	0.932* (0.898–0.955)	243/307	0.792 (0.743–0.833)

*Subjects helped define the reference standard of clinically important relations.

system only captured part of the temporal information, its answer was judged incorrect. For example, a patient developed a rash one week before admission, but the system only captured "before admission."

Compared with the reference standard, the temporal reasoning system incorrectly answered 16 questions. In addition, the system could not answer 8 questions since the medical events were not extracted by MedLEE. For example, terms like "rheumatological consultation," "GI button (gastrointestinal button)," and "declared" in "the patient was declared" were not extracted by MedLEE. Therefore, the overall accuracy of the system in answering temporal queries was 83.7% (123 out of 147; CI: 76.9–88.8).

We further ascertained the causes of the errors. Among 16 incorrect answers, four answers provided incomplete information. For example, for the statement, "well until one week ago when she developed papular rash on the neck," the system did not link one week ago to rash, but only inferred "before admission." The system is not designed to handle age information at this stage, so that for sentences like "the patient was diagnosed with cystic fibrosis at age four," the system only inferred "the diagnosis of cystic fibrosis was made before admission" but not the exact year when the diagnosis was made. The system misinterpreted some expressions. For example, the system misinterpreted "on 1/2" in "the patient was put on 1/2 maintenance IV fluids" as a date. Misplaced contents (e.g., the statements about hospital course were misplaced in the section of "history of present illness") caused the systems use inappropriate rules in the knowledge-based subsystem. For example, as noted above, one report had no hospital course section. All the information was in the history of illness, physical examination and laboratory test sections. Therefore, questions like "did the patient use heparin during hospitalization" could not be answered properly.

To get the right answer, complex queries are necessary for some questions. Manual checking was used to assist in finding the answers. For example, a term, "Bactrim," appeared several times in a report. If we want to know "was the patient treated with Bactrim during hospitalization," a manual summarization of retrieved temporal information about all the occurrences of "Bactrim" is needed.

Discussion

We found that the TimeText system generated many temporal relations, that most of them were correct (97%), and that it generated most of the temporal relations deemed clinically important by subjects and raters (79%). The human subjects achieved a similar level of correctness. They captured a higher proportion of the clinically important relations, but they helped to create the reference standard. When the relations were placed in a database and queried, the system answered 84% of 147 time-oriented questions correctly. This compared to 88% correct for the experts when compared to each other.

This study is one of the few attempts in the literature to assess temporal reasoning systems for medical text. It is difficult to evaluate a system that processes medical narrative data:[23,25] 1) it involves much manual processing by domain experts; 2) inter-rater and intra-rater agreement may be low; and 3) obtaining a gold standard is difficult. In addition, temporal reasoning using medical narrative data involves complex reasoning and calculations, which places an even heavier burden on the experts.

Hirschman et al.[13,26] developed "the time program" for obtaining a representation of time for each medical event stated in a discharge summary, either in terms of a fixed time point, or in terms of another events in the narrative. They also applied a special time comparison retrieval routine which compared the temporal information for two events and returned one of four values: greater than, less than, equal, or not comparable. Only three discharge summaries were used to assess the performance of the system on retrieving clinical information. The system-generated responses showed 90% agreement with the results obtained by a physician reviewer. However, their evaluation methods were not described in detail.

A report by Rao and colleagues[15] described a system, called REMIND, for inferring disease state sequences for recurrence using both clinical text and structured data. Phrase spotting was applied to information extraction from free text and a Bayesian Network was used for temporal inference. They assessed REMIND's classification accuracy (whether the patient recurred or not) and sequence accuracy (if the patient recurred, did the system correctly estimate the disease-free survival time). The purpose of this study differed from ours in that they focused on specific recurrent medical events instead of different events. Bramsen et al.[16] described a supervised machine-learning approach for temporally segmenting discharge summaries and ordering these segments. They defined a temporal segment to be a fragment of text that does not exhibit abrupt changes in temporal focus. Their learning method achieved 83% F-measure in temporal segmentation, and 78.3% accuracy in inferring pairwise temporal relations. Compared with this approach, the TimeText system performs temporal analysis at a finer granularity.

The TimeText system generates the timelines from three sources: 1) the constraints encoded in the temporal constraint structures, which represent only what is stated explicitly in the report; 2) the constraints discovered using linguistic and medical domain knowledge, which include implicit information; and 3) the constraints derived from resolving the simple temporal constraint satisfaction problems, which include derived information. Compared with the system, the human subjects tended to focus on listing temporal relations for the events that occurred next to each other in a timeline. They mentioned that transitive relations can be inferred based on this information but that they might not list the inferred relations unless they were very important. As the result of using different strategies for timeline generation, TimeText generated three times more temporal relations than the annotators. Our belief is that many of these additional relations are obvious to humans, and so they do not bother to write them down. Our system infers these relations ahead of time, but they could in theory be generated by a reasoning system in the process of answering a question.

While many challenges exist specifically for the system, some difficulties are common both for the physicians and the system. We found that most of the temporal assertions