G Model
ARTMED-1546; No. of Pages 14
ARTICLE IN PRESS

Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain

Ayelet Goldstein[a,*], Yuval Shahar[a], Efrat Orenbuch[b], Matan J. Cohen[c]

[a] Medical Informatics Research Center, Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[b] Hadassah Medical Center, Jerusalem, Israel
[c] Clalit Health Services, affiliated with the Hebrew University School of Medicine, Jerusalem, Israel

ARTICLE INFO

Article history:
Received 30 October 2016
Received in revised form 16 August 2017
Accepted 1 September 2017

Keywords:
Medical informatics
Natural Language Generation
Textual summarization
Temporal abstraction
Quantitative evaluation
ICU

ABSTRACT

Objectives: To examine the feasibility of the automated creation of meaningful free-text summaries of longitudinal clinical records, using a new general methodology that we had recently developed; and to assess the potential benefits to the clinical decision-making process of using such a method to generate draft letters that can be further manually enhanced by clinicians.

Methods: We had previously developed a system, CliniText (CTXT), for automated summarization in free text of longitudinal medical records, using a clinical knowledge base. In the current study, we created an Intensive Care Unit (ICU) clinical knowledge base, assisted by two ICU clinical experts in an academic tertiary hospital. The CTXT system generated free-text summary letters from the data of 31 different patients, which were compared to the respective original physician-composed discharge letters. The main evaluation measures were (1) relative *completeness*, quantifying the data items missed by one of the letters but included by the other, and their importance; (2) *quality* parameters, such as readability; (3) *functional* performance, assessed by the *time needed*, by three clinicians reading each of the summaries, to answer five key questions, based on the discharge letter (e.g., "What are the patient's current respiratory requirements?"), and by the *correctness* of the clinicians' answers.

Results: Completeness: In 13/31 (42%) of the letters the number of important items missed in the CTXT-generated letter was actually less than or equal to the number of important items missed by the MD-composed letter. In each of the MD-composed letters, at least two important items that were mentioned by the CTXT system were missed (a mean of 7.2 ± 5.74).

In addition, the standard deviation in the number of missed items in the MD letters (STD = 15.4) was much higher than the standard deviation in the CTXT-generated letters (STD = 5.3).

Quality: The MD-composed letters obtained a significantly better grade in three out of four measured parameters. However, the standard variation in the quality of the MD-composed letters was much greater than the standard variation in the quality of the CTXT-generated letters (STD = 6.25 vs. STD = 2.57, respectively).

Functional evaluation: The clinicians answered the five questions on average 40% faster ($p < 0.001$) when using the CTXT-generated letters than when using the MD-composed letters. In four out of the five questions the clinicians' correctness was equal to or significantly better ($p < 0.005$) when using the CTXT-generated letters than when using the MD-composed letters.

* Corresponding author at: Medical Informatics Research Center, Department of Information Systems Engineering, Ben-Gurion University of the Negev, P.O.B. 653, Beer-Sheva 8410501, Israel.
E-mail address: gayelet@bgu.ac.il (A. Goldstein).

http://dx.doi.org/10.1016/j.artmed.2017.09.001
0933-3657/

Conclusions: An automatic knowledge-based summarization system, such as the CTXT system, has the capability to model complex clinical domains, such as the ICU, and to support interpretation and summarization tasks such as the creation of a discharge summary letter. Based on the results, we suggest that the use of such systems could potentially enhance the standardization of the letters, significantly increase their completeness, and reduce the time to write the discharge summary. The results also suggest that using the resultant structured letters might reduce the decision time, and enhance the decision quality, of decisions made by other clinicians.

1. Introduction

1.1. The need for interpretation and summarization of time-oriented clinical data

Many clinical tasks require dealing with very large amounts of patient data. Physicians, who have to make diagnostic or therapeutic decisions, may be inundated by the volume of data if their ability to reason does not scale up to the amount of data.

The need to interpret and abstract large amounts of data is especially crucial in the case of patients who have chronic diseases. In such cases, the physician often needs a brief free-text summary of the patient's course over the past several months or years. An effective summary often requires the creation of several levels of abstraction, from raw data to high-level interpretations. Thus, the interpretation and summarization of clinical data, especially in the case of chronic patients, consumes large amounts of human time and effort. However, interpretation and summarization of large numbers of clinical data, even over relatively short spans of time, such as in the *Intensive Care Unit* (ICU) context, is quite a challenging task as well.

1.2. The importance of discharge summaries

A patient *handoff* is the transfer of care from one care provider to the next [1]. The handoff serves a critical function in hospitals because it is a core process in ensuring patient care continuity. Handoffs occur in times of transition, in which clinicians share and process patient information as well as jointly plan the next steps of patient care. Receiving care in multiple settings means that patients receive medications from different clinicians, presenting increased opportunities for medication errors to occur. Due to its importance during a patient's hospitalization, a poor handoff is linked to adverse effects and poor outcomes, including delays in treatment and ordering of tests, and increased length of stay [2–6].

Given the importance of the handoff process, we decided to focus in the current study on the automated generation of a free-text discharge summary from the structured quantitative, multivariate, longitudinal data of a patient's record, since a textual discharge summary is the most common method of handing-off patients from one unit to another. In particular, we opted to focus on a discharge summary of patients who were hospitalized in the ICU and are being transferred to another ward, since the handoff from the intensive care unit to the medical ward poses a number of unique risks to patients recovering from critical illness.

1.3. The CliniText framework for knowledge-based textual summarization of longitudinal clinical records

There are a variety of approaches to perform *text-to-text* summarization [7], namely, abstracting a corpus of text as another (shorter) text document; some of these approaches are statistical, some employ natural-language processing techniques, some focus on machine learning methods, and some use hybrid techniques [8]. Typically, these methods either extract actual key sentences from the text that represent the full text, or generate a summary that is semantically equivalent to the original text. However, performance of a *data-to-text* summarization, which is our objective here, requires completely different techniques.

To perform an automatic summarization of longitudinal clinical records that include time-oriented data of different types and from multiple sources (i.e., multivariate time-oriented clinical data), we had previously developed and implemented a new framework for natural-language generation that exploits clinical domain knowledge. Given time-stamped numeric and symbolic raw clinical data (which include measurements such as temperature, SpO2; drugs received; procedures performed such as blood transfusion, physiotherapy), and a suitable domain-specific knowledge base, our new process generates a textual summary of the data.

We had implemented the new framework for text generation in a new software system, which we refer to as the CliniText (CTXT) system. The approach taken by the system is generic and is not specific to any clinical domain [and not even limited to generating only discharge letters-other types of summaries are possible as well]; we had demonstrated the feasibility of the CTXT system to produce general summaries, although only in a preliminary fashion, also in the cardio-thoracic ICU and in the diabetes domains [9,10].

The CTXT system first abstracts the raw time-stamped clinical data in the patient's record into higher level, meaningful, interval-based concepts, using the *knowledge-based temporal-abstraction* (KBTA) method [11,12]. The KBTA-based temporal-abstraction system underlying the CTXT system has been evaluated in several previous studies, in several different clinical domains, and in the context of several different tasks – from data summarization and visualization, to monitoring and therapy [12–15], and had been demonstrated to be quite sound (although not necessarily complete, with respect to all of the clinically meaningful temporal patterns that can be extracted from the raw, time-stamped data).

The CTXT system then adds missing events, such as implicitly assumed clinical procedures, through an *abductive reasoning* process. The resulting set of raw and abstract data is *pruned* using several heuristics, some of which refer to the domain's temporal-abstraction ontology. Then, the pruned concepts undergo *document structuring*, using a pre-defined document tree, to define the order in which the concepts will appear. A *microplanning* module groups the concepts at several levels of abstraction, such as by context or by value. Finally, the only language-dependent module, the *surface-realization* module, generates the actual natural-language text. A full description of the CTXT system's architecture and computational processes and of its several innovative features is outside of the scope of the current paper, and can be found elsewhere [16]. However, a brief description of its architecture and computational processes is presented in Fig. A1 in the Appendix A.

The CTXT system essentially performs a *Natural Language Generation* (*NLG*) task. This task has previously been performed in several domains. However, in most of the implemented systems, the character of the data to be presented was relatively well-defined [17,18]. As a result, although sophisticated NLG technology for linguistic