On Regular Analysis of Medical Reports

Jiri Semecky, Jana Zvarova
European Center for Medical Informatics, Statistics and Epidemiology – Cardio,
Institute of Computer Science AS CR, Pod Vodarenskou vezi 2, Prague, The Czech Republic

Abstract:

The most of the Czech software systems for acquiring and storing patient data use free–text patient records. However, structured form of data supports automatic processing, e.g. statistical evaluation or decision support. The goal of this work was to analyze free–text medical reports to find information that could be semantically determined and described, and so stored as structured data. We used 120 anonymized free–text medical reports of cardiac patients from two Czech hospitals. The analysis used regular grammars described by rules built up from 20 training medical reports. We used the cardiology knowledge base created at the EuroMISE Center–Cardio for building the rules. Other 100 records were used for testing purpose.

INTRODUCTION

The importance of electronic patient record in cardiology (EPR) for data acquisition, data storage and data mining is considered. We have reviewed different databases from the field of cardiology. Our goal was to establish a cardiology database that can be used to compare practice with guidelines and discover some features of diseases that can help to their management. Therefore medical knowledge in different medical guidelines can be repeatedly evaluated using large amount of data collected by EPRs and stored in a cardiology database.

Data collected in cardiology have different quality and accuracy. Therefore minimal data model for cardiology has been developed to assure minimal data collection with high quality and accuracy over a long period of time and further improvement of cardiovascular diseases management. Minimal data model for cardiology covers three following areas. First, it is necessary to create a list of information items to be collected. Second, measurement scales for selected information items are chosen. Third, a logical relationship of items needs to be defined.

Principles on which the data model is based are objectiveness, reliability, validity, comparability, economy and completeness [1]. These principles are very important for data model of any complex disease. *Objectiveness* assures that data are independent of the collecting clinician. *Reliability* means that data are reproducible with a high accuracy from other experts of the field. *Validity* means that collected data are describing what should be described. *Completeness* reflects the completeness of relevant information. *Comparability* assures that diagnostic methods and examinations proving the same facts lead to the same results. *Economy* performed and activities (most often in the form of drug administration) are carried out.

The large database in cardiology, based on appropriate data model, can serve as a source of information for decision support systems and can be explored by data mining methods to reveal hidden associations and relationships. Moreover, EPR serving as a tool for data collection can be used for automated generation of alerts, reminders and suggestions when standards of care (e.g. based on medical guidelines) are not achieved. There are different approaches how to improve the decision support in this field. However, still the big problems occur with handling of computerized patient records (CPR), while the most of information is stored in the textual form. In this context we will focus on the application of regular semantic analysis of the text medical reports that can lead to structured form of the information.

SFT SYSTEM

In the paper we deal with the important problem of mutual relationship between the structured form and free–text form of the electronic patient record. In these days Czech health information systems use only free–text computerized patient records; it

means storing the most of information as a free–text. In the work [2] a new system SFT (Semi Free Text) has been proposed. It can store patient data in both forms, structured form and free–text at once. The SFT system implementations are based on the implementation of the knowledge base that defines the hierarchical structure of information on patients. The data are stored in the database in a form that allows both views, free–text and structured form view. The part of the SFT system is also the module for the automatic analysis of the free text using regular grammars. The SFT system was motivated by the European project MGT [5] and by the experience gathered in the European project TripleC [3], where the ORCA (Open Record for Care) system was tested and adapted to the Czech language environment [4]. In the MGT project textual information of paper medical guidelines have been transferred into electronic medical guidelines.

In this paper we judge the possibilities of automatic semantic analysis of patient medical reports. The semantic analysis is connected to the medical knowledge base, which is used to structure the electronic health record. Therefore the semantic analysis of a medical report is an algorithm searching for fragments of the medical report with the information described by the given knowledge base. The first possibility of semantic marking is the application of linguistic methods of the text analysis. These methods cover both lexical and syntactical text analysis. However, in medical reports written in the Czech language, the most of the information is covered in short textual messages that do not have structure of the Czech sentences. For this reason the method of semantic marking using regular grammars is proposed.

It is a big advantage that this method is not bounded with the concrete language. Moreover, it is not so much time consuming and there is no need of other large data sources, e.g. vocabularies. Therefore the semantic marking of medical reports is an algorithm that provides semantic analysis according to given rules and marks (e.g. using XML tags) the founded fragments of the medical report with a reference to the given knowledge base.

For the implementation of the SFT system the three–layer architecture was used. The application layer is implemented as a PHP-module that runs on the Apache web server. The presentation layer consists of two modules. First, the SFT-client that allows viewing and editing SFT electronic health records. Second, the regparser client that pursue the regular analysis itself. The regparser client is implemented in the JAVA language. Borland InterBase server presents the data layer.

The knowledge base is the ground of the semantic marking and it's stored on the database server. In [2], the way in which the knowledge base was constructed is described and there is designed the XML-based language KBML (Knowledge Base Markup Language) describing the knowledge base.

For the communication between the application and the presentation layer an XML-based language SFTQML (Semi Free Text Query Markup Language) is designed. The requests of the application layer are usually simple commands, whereas the answers are large structured texts.

TESTING OF THE REGULAR ANALYSIS IN CARDIOLOGY

We have tested the regular analysis on anonymized medical reports from two Czech hospitals (University Hospital in Prague and Municipal Hospital in Caslav) participating in the research center EuroMISE-Cardio. The SFT system used the medical knowledge base developed for the minimal data model in cardiology at the EuroMISE Center – Cardio [6].

The regular analysis was trained on 20 medical reports from the Municipal hospital in Caslav (MHC). We have searched these reports for information that could be described by our knowledge base and we have constructed rules

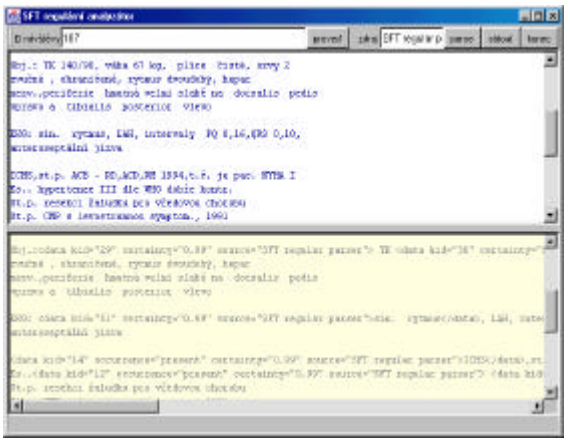

Figure 1: The regparser program

according the found appearances. After finishing that process we have run the automatic analysis on the training set and corrected the rules according to the mistakes being done. This process let us to 60 semantic rules, each one representing a regular grammar.

The automatic semantic analysis itself tests the text (medical report) on each rule and marks founded fragments of the text with a reference to the knowledge base.

Example: Let's consider we have a rule that could be represented as
BP ((<int>)$_{\text{systolic blood pressure}}$ /(<int>)$_{\text{diastolic blood pressure}}$)$_{\text{blood pressure}}$.
where <int> represents an integer value and parenthesis with an index represent the reference to the knowledge base. Now let us consider, there is an occurrence of the text "BP 120/80" in the analyzed medical report. The semantic analyzer will be mark "120/80" as a blood pressure value (by pointing to the knowledge base) and the values 120 and 80 will be marked as systolic blood pressure and diastolic blood pressure.

In the SFT system the marked text would look
BP <data kid="5"><data kid="6">120</data>
</data>/<data kid="5"><data kid="7">80</data>
</data>
where 5, 6, 7 are pointers to corresponding items in the knowledge base.

The semantic marking of medical reports was done by the program regparser. Figure 1 shows a result of the regparser program running on a Czech medical report.

Table 1. Results of regular analysis

Variable	training set of MHC	testing set of MHC	testing set of UHP
No. of medical reports	20	100	30
Size of text [B]	28 082	182 976	64 663
No. of semantic values	159	1122	231
No. of semantic values per message	7.95	11.22	7.70
No. of semantic values per kB	5.66	6.13	3.57

As the training set we used 20 medical reports from the Municipal hospital in Caslav (MHC). Based on these medical reports and the cardiology knowledge base the rules of regular analysis were developed. As the testing data set we used 100 of medical reports of the same hospital and 30 medical reports of the University hospital in Prague (UHP).

The Table 1 shows the success of the regular analysis measured by the number of semantic values found in one medical report and in one kB of text. In the case of the University hospital in Prague, the number of found semantic values per one kB of text was lower then for medical reports from the Municipal hospital in Caslav. It could be caused by different way of recording medical information by clinicians in different hospitals or by different syntactic constructions of medical records.

Finally we have measured the precision and the recall coefficients [7] of the analysis on 10 medical reports from each hospital. The precision coefficient (P) is defined as the ratio of the number of correctly found semantic values to the total number of found semantic values (including correct marked even wrong marked ones). The recall coefficient R is defined as the ration of the number of correctly found semantic values to the number of all relevant values (that should have been marked). The Figure 2 and Figure 3 show the outcome of our measurement on the medical reports from the Municipal hospital in Caslav and from the University hospital in Prague.

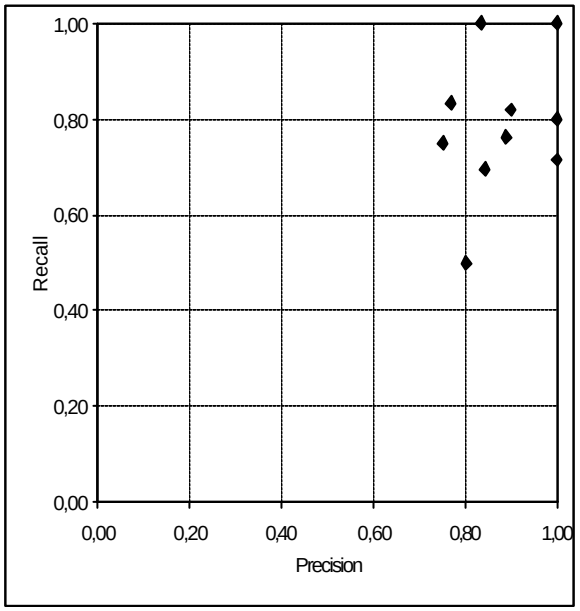

Figure 2. Results of regular analysis for 10 medical reports of the Municipal hospital in Caslav

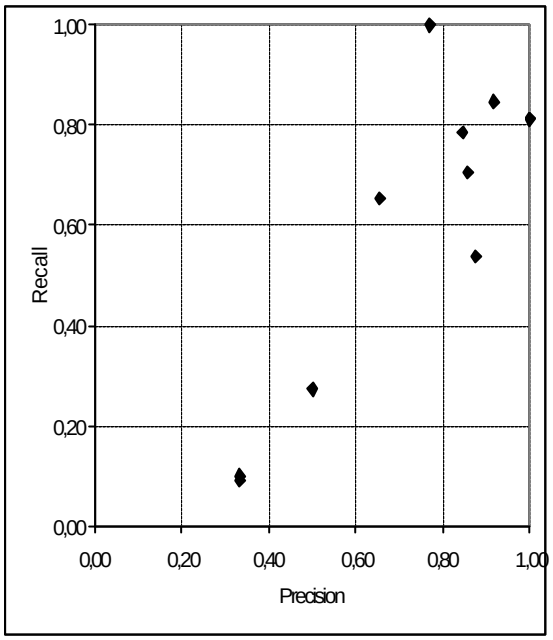

Figure 3. Results of regular analysis for 10 medical reports of the University hospital in Prague

CONCLUSION

The outcomes of our research implies that the rules constructed on medical reports from one hospital are successful to a certain extend for analyzing the reports from the same hospital, while they're slightly worse for the reports from another one. It is caused mainly because of another way of writing text. If we used medical reports from both hospitals as a training set, the result would be probably more harmonious.

Semantic analyses of medical report is not used anywhere in the Czech republic. Yet, its usage could be beneficial for data analyzing or for searching. There are limitations to the regular analysis, because the regular languages are not strong enough to characterize more complex sentence constructions. It would be probably useful to employ the natural language lemmatization and syntax parsing here. Still there would be some problems, while medical reports rarely have a form of grammatically correct sentences.

ACKNOWLEDGEMENT

Research was partially supported with the project LN00B107 of the Ministry of Education of the Czech Republic.

REFERENCES

1. Mannsmann U, Taylor W, Porter P, Bernarding J, Jager HR, Lasjaunias P, TerBrugge K, Meisel J. Concepts and Data Model for a Co-Operative Neurovascular Database. Acta Neurochirurgica 2001;143: 783-791
2. Semecky J, Multimedia electronic patient record in cardiology. Diploma thesis supervised by J. Zvarova. Charles University, Prague 2001
3. Pribik V, Grunfeldova H, Hanzlicek P, Peleska J, Zvarova J, Czech national data standards implementations in ORCA electronic patient record in cardiology. Medical Infobahn for Europe, A.Hasman et al. (eds.), IOS Press, Amsterdam , 2000; 652-655
4. Pierik FH, van Ginneken AM, Timmers T, Stam H, Weber RF. Restructuring routinely collected patient data: ORCA applied to andrology. Methods of Information in Medicine 1997; 36:184-190.
5. Svátek V, Kroupa T, Ružicka M. Guide-X - a Step-by-step, Markup-Based Approach to Guideline Formalisation. Computer-Based Support for Clinical Guidelines and Protocols , Studies in Health Technology and Informatics 83. B. Heller , M. Löffler, M. Musen and M. Stefanelli (eds.), IOS Press, Amsterdam , 2001
6. Zvárová J. Clinical Databases Originating in Electronic Patient Records. Biocybernetics and Biomedical Engineering. 2002; 22: 43-61
7. Su LT. The relevance of recall and precision in user evaluation. Journal of the American Society for Information Science 1994; 45:207-217