

Data Processing from mHealth Patient Data Acquisition Related to Extracting Structured Data from EH Records

Stefan Balogh, Fedor Lehocki, Daniel Ivaniš, Erik Kučera,
Miloš Lajtman, and Igor Miňo

Slovak University of Technology, Faculty of Electrical Engineering and Information
Technology, Institute of Computer Science and Mathematics, Ilkovicova 3,
812 19 Bratislava, Slovakia
{stefan.balogh, fedor.lehocki, daniel.ivanis, erik.kucera,
milos.lajtman, igor.mino}@stuba.sk

Abstract. Application of mobile devices in healthcare is a pervasive way how to assess patient health status. Adding just another data source to overwhelmed physician requires technologies for their effective processing. Despite the fact that the problem of extracting clinical information from free-text health reports for computerized applications or for decision support systems is a lot discussed issue, it is still complicated and unresolved question. In general Natural language processing (NLP) systems are implemented to solve the task. However NLP system works only for relatively narrow clinical domains because the format of the language used in the reports is not standardized and the reports vary depending on the domain. We describe methodology suggested for extracting structured data from EHR focusing especially on EHR in Slovak language. Further, we discuss problems concerning a task of extracting required data from free-text health reports. In the conclusion we present test results and possible implementations.

Keywords: Natural language processing, NLP, health reports, structured data, telemedicine.

1 Introduction

Increasing healthcare costs and shortage of physicians pave the road for new technologies to enter the healthcare sector. Although mHealth is around for some time, wide spread of mobile WWAN/ WLAN networks and cost effective wireless sensor equipment enable real pervasiveness of mobile health. This provides reliable and cost effective access for patients and physicians to health monitoring, data collection, delivery and support of health information, diagnosis and treatments, research and education. In this paper we would like to emphasize more on what happens “behind the curtain” i.e. when data are collected and stored. We believe that adding another data source to fill up data storage will not improve the quality of information or bring knowledge to support physician in his daily routine. Not all data collected by mHealth solution are structured. For example in case of self-management of diabetes [1], besides creating the structured data related to glycemia measurements and other relevant data,

patient can also create custom notes of unstructured text. These notes include observations about patient's general health with specific symptoms (in case of infection) or comments related to diabetes (hypoglycemia during the night). It would be interesting to incorporate methods for analyses of free-text in electronic health records (EHR) as part of mHealth solution related to data processing [2]. Data from EHR can be used also for provision of clinical decision support services. The unstructured content of patient's reports stored in EHR represents a rich source of data. Therefore it becomes a major bottleneck hindering widespread deployment of effective clinical applications because it is difficult, if not impossible, to access the textual information reliably by computerized processes [3]. For solving this problem different Natural language processing (NLP) systems were developed. A NLP-based approach has been applied for a variety of problems. These include identifying patient cohorts, reporting of diseases, syndrome surveillance, diagnostic classification, identifying co-morbidities, medication event extraction, adverse event detection, identification of postoperative complications, and disease management [8].

The paper focuses on describing the problems that are related to the processing of unstructured text contained in medical records taking into account various possible forms of the records defined for example by particular writing style of a physician. Aim of the presented work is to highlight design of the whole process of obtaining structured information from unstructured documents. Being aware of the particularities of Slovak language, we tried to describe all particular phases of the whole extracting process in detail. Sections 1.1 and 1.2 describe the existing work and specific problems, in section 2 we focus on suggested methodology, and in the section 3 related to conclusion we discuss the advances and shortcomings of the presented methodology.

1.1 Related Works

The related works can be divided into those where the text parser for concrete task in clinical area is described and those which deal with the text parser for general solution searching approaches. Friedman [3] analyzes semantic methods that map narrative patient information to a structured coded form. In [4] challenges are discussed related to processing of clinical reports like performance, availability and confidentiality of clinical texts, intra- and inter-operability. Other works concentrate more on development of accurate decision support system and free-text processor for extracting required information from reports. For example Aronsky *et al* [5] has demonstrated the development of accurate clinical decision support system (CDSS) for diagnosing pneumonia that utilized a free-text processor for pneumonia reports. Demner-Fushman *et al* [6] have discussed the potential of NLP to enhance CDSS. Dupuis *et al* [7] recently constructed a free-text parser to identify abnormal Pap reports. In [8] authors claim that the application of NLP for clinical decision support has been deferred, possibly because of the requirement for a high accuracy.

1.2 Specific Problems in NLP Systems

The main known problems which occur during application of NLP systems for extracting structures and codification of clinical information from the health reports

are already described in [3]. It includes expressiveness, heterogeneous formats, abbreviated text, interpreting of clinical information, rare events. Also, physicians use their own special terms and abbreviations, so each report can have different terms, depending on an author (physician). Taking into account these problems we would like to present a design of a theoretical model for extraction of clinical structured information for medical application. Our target is to map the extracted information for a controlled vocabulary and for a standard domain representational model. The representational model of medical language is essential for interpreting the underlying meaning of clinical information in the reports and relations among the information. Creating such model is a demanding task already. Moreover, mapping of extracted terms from the health reports for that model is even more challenging.

2 Methodology

Core NLP system usually uses syntax and semantics analysis for achieving the correct result. Syntactic analysis determines the structure of a sentence and the relationships among the words in the sentence. Semantic analysis is a process that determines what words and phrases in the text are clinically relevant, and determines their semantic relations. An important requirement of semantic analysis is a semantic model of the domain or ontology [3]. The NLP system usually has components for morphological analysis, lexical look up, syntactic analysis, semantic analysis and encoding, still many variations are possible. More about these components can be found in [9]. For our purpose of extracting clinical information from various kinds of reports we have created our own methodology which deals with problems mentioned in previous section. The methodology consists of two phases, the preparation phase and the processing phase.

The preparation phase has the following activities: 1. Morphological analysis – (lemmatization); 2. Creating a list of words used in reports; 3. Lemmatization completion; 4. Manual division of reports into logical sections; 5. Training the machine learning algorithm; 6. Obtaining required information from logical section into defined structure.

The processing phase consists of the following activities: 1. Lemmatization of a report; 2. Division of report into sections; 3. Encoding the required information to defined structure.

Further we describe all the activities in more detail for better understanding.

2.1 Preparation Phase

Preparation phase is necessary to avoid the previously mentioned problems for NLP systems and to adapt the system to real situation. It can be achieved by text normalization in reports. The medical reports are frequently written in technical language using loan words and their derivations, even in neologisms. Some

physicians use their own special terms or abbreviations. Therefore, it is required to figure out a way to include these words in the process of lemmatization or, in case of abbreviations, to translate the words into their original form. To break up the original words in a text to their canonical forms in Slovak language we used lematizator called Morphonary [12]. Morphonary works with three dictionaries - Dictionary of Foreign/Derived Words (Slovník cudzích slov- SCS), Dictionary of the Slovak Language (Slovník slovenského jazyka- SSJ) and the Declined Words Dictionary. SCS contains about 60,000 words in the base form, SSJ two times more. The key vocabulary in this methodology is the Declined Words (DW) Dictionary. This dictionary contains 1730 words in the base form, as well as their all declined (inflected) forms. This dictionary is composed from a selection of such terms and words that make the best representation within variability of declined forms of words. In this dictionary there are pairs of words "basic form - declined form". During preparation phase we configured all components for reports from department of internal medicine taking into account text formats, sections format and specific words.

In the preparation phase we make **lemmatization** [12] of chosen number of reports using Morphonary and DW dictionary with default words. It is required in the process of examination of the list of the most common words in the reports. In the list we can find special medical terms or physicians own custom expressions, abbreviated text and words not included in our dictionary and therefore not corrected during lemmatization. In cooperation with medical staff we tried to find the real meaning of the words or abbreviated text. Then, during the lemmatization completion activity, we created a new pair in the Declined Words Dictionary. For example, if doctors used the abbreviation "pcnt" rather than the "patient", we have added to the DW dictionary pair "pcnt - patient". Then during lemmatization process this atypical shortcut is replaced with usual word "patient". Also we divided the report into logical sections manually in a chosen number of reports to enable training of the machine learning algorithm. As a last activity in this phase we choose an algorithm for obtaining required information from report into defined structure.

Creating a list of words used in reports - We examined the lemmatized texts using the application RapidMiner¹ and have created a list of the most common words. To obtain a list of words (wordlist) of selected documents we have used procedure including pre-processing, text tokenization and removing stop-words. Wordlist conveys the desired list of words. Therefore we removed the common words with the aid of dictionary of stop words. This provides the list of words and abbreviations that are specific to the reports and they should be identified and included into the DW dictionary.

Manual division of reports into logical sections - Logical section can be understood as part of a report describing the relevant topic, e.g. anamnesis, medications, laboratory results, diagnosis, etc. It can consist of several paragraphs (or vice versa).

Division of a report to its logical sections is necessary to provide correct data for machine learning algorithm. Usually logical section of a report can be generally

¹ <http://rapid-i.com/content/view/181/190/>

recognized but sometimes it must be discussed with the specialist who created the report. The first step is to split the text into paragraphs according to the written arrangement (usually defined by the writing style of a physician) of a document. It is based on the assumption that physician is also dividing text into paragraphs during creation of the patient's report.

During analysis of the available documents (approximately 3000 reports), in many cases text was structured into different paragraphs, although the unit has the same semantic part (logical section). To structure the parts of the texts in such documents doctor leaves out a line for each paragraph. Based on these findings, we implemented in the application the possibility of dividing a document according to paragraphs or left out lines. For the case of our reports we have divided paragraphs into nine sections: 1. Doctors' names, outpatient facilities; 2. Objective examinations and laboratory testing; 3. Medications, therapy; 4. Recommendations, conclusions; 5. Epicrisis; 6. ID code of physician and specialization; 7. Subjective problems; 8. Meaningless characters; 9. mixture of various categories.

Training the machine learning algorithm - This training is essential for the algorithm in categorization while performing supervised learning. The task is to train the algorithm for each kind of reports for its specific logical sections. Therefore, for the purposes of training data, only manually divided logical sections from the previous activity were used. For logical structure of our reports (divided into 9 sections as mentioned above), the most suitable was naive Bayes classifier.

Obtaining required information from logical section into defined structure - This part of the process requires defining the information and terms from the logical sections of the individual medical records and finding suitable methods for their accurate identification. It is usually not easy to decide on the right approach for solving this type of task. As we mentioned above we can use syntax tools or tools based on semantic analysis. But semantic approaches, while more precise, are subject to poorer coverage than syntax approaches [10]. Therefore in many cases the balanced approaches are used. Balanced approaches utilize more equal amounts of syntax and semantic processing. First comes the syntax processing and then the semantic analysis is applied to eliminate incorrect syntactic parse trees and to further identify domain words. Despite measures such as dividing into section it is necessary to use more advanced methods for correct identification of expression. The rule base may vary for each term and it is not possible to establish a general procedure. Therefore solutions need to be found for each term individually. The next necessary step is setting up the system that receives and manages to save those terms. In some cases, to correctly recognize them it is essential to use semantic analysis and create ontology. For example, in our case we configured the system for identification of drugs, blood pressure, and diagnosis. More complex structures for drugs were created that contained the name, dosage and prescribed quantity. Our rule base for obtaining structured information was developed and tested using the corpus of 3 000 test reports from department of internal medicine. We can see the example of identified structure related to drug data in report (Fig. 1) and the identified and filled structure of blood pressure on Fig. 2.

Type	Set	Start	End	Id	Features
Lieky		30	48	4222	{rule=Lieky}
Lieky		50	61	4223	{rule=Lieky}
Lieky		63	78	4224	{rule=Lieky}
Lieky		80	92	4225	{rule=Lieky}
Lieky		94	114	4226	{rule=Lieky}

Fig. 1. Identified the structure of drugs

2.2 The Processing Phase

The processing phase has following activities: lemmatization of report, division of report into sections and encoding a particular phrase to relevant structure.

Lemmatization of report - For lemmatization we have used the DW dictionary prepared in preparation phase, so we can eliminate the problem of ambiguous abbreviations and specific medical terms in the given type of reports.

Division of the report into sections - For division we have used the chosen machine learning algorithm from preparation phase. The output is stored in text files which are systematically labeled by index number of the paragraphs in the original document. This ensures the possibility to re-stack the original document. Thanks to this extension we have solved the problem of heterogeneous report format and also this helps us to deal better with expressiveness-and interpreting clinical information.

KrvnyTlak

R diastolický 80

R systolický 120

C

TK:120/80torr

Open Search & Annotate tool

Fig. 2. Filled structure for blood pressure

Encoding the required information to defined structure - This was the last activity, where we have searched for specified information and filled them into desired structure for each logical section separately. We have used rule bases created in the preparation phase for processing of each logical section. Example of encoded particular phrase to a relevant structure is shown on Fig. 2.

3 Conclusion

The whole process is more complex in the final realization and it depends on many conditions. The main distinguishing factor is the purpose for which the information is gained, whether it is a single-purposed application based on a decision support system or whether the information must be available for different applications. In the latter case it is important to achieve the maximal re usability of data with respect to existing standards (e.g. archetypes). Importance of the whole methodology does not lie in the design of the process phase, but in the design of methodology of the preparation phase. The process stage is very simple. Lematize report and divide it into sections using selected and pre-configured techniques to store information to designated structures which can be used by other applications and systems. In the preparation phase the primary task is the setting up of lematizator. The new idea is using lemmatization process and Morphonary DW dictionary for Slovak language to translate and map the unknown words and the clinically relevant terms to well defined concepts in a controlled vocabulary (such as the known clinical healthcare terminology vocabularies like UMLS, ICD-9 or SNOMED) and to the original form of words. Also, new approach is represented in use of wordlist and stop words dictionary to filter familiar words which support easier identification of special unknown terms, completion of abbreviations and terms unknown to lematizator. The setting of the method for dividing paragraphs into given optional sections will extend to exploration and testing of different methods for text categorization and testing their effectiveness for the type of records. For finding and correct recognition of clinical information and inserting a particular phrase to a relevant structure (e.g. standard archetypes), syntactic and semantic analysis is used. The creation of such systems is also not a trivial problem. At the same time the tools that use syntax parsing seek to relate semantically relevant phrases via the syntactic structure of the sentence. In this sense, syntax serves as a "bridge" to semantics [11]. So we must find combination of syntax tool with suitable semantic one to achieve the goal. Cooperation with medical personnel is critical in achieving this goal. Without their support it would not be possible to configure the system or to create the ontology correctly (not mentioning the identification of abbreviation or special words). The methodology we have created had some shortcomings related to elegance in its application in different phases and activities (e.g. lematization, search and encoding information to relevant structures, splitting the unstructured text into logical sections, etc.). This also includes selection of a suitable machine learning algorithm for division report into logical sections. Overall, presented methodology provides complete and complex solution with explicitly defined phases and activities. Each activity can be implemented with

different approach than described above. This is determined by the logical structure of the unstructured patient report and type of information that we search for in the text, i.e. position of desired information in the text like drugs, diagnosis, lab results etc.

Acknowledgments. The authors would like to acknowledge the support for research under following projects: Measuring, Communication and Information Systems for Monitoring of Cardiovascular Risk in Hypertension Patients (APVV-0513-10); Analytics Services for SMARTer Healthcare (IBM SUR project); Research & Development Operational Programme for the project Support of Center of Excellence for Smart Technologies, Systems and Services I and II (ITMS 26240120005, ITMS 26240120029), Competence Centre of Intelligent Technologies for Electronisation and Informatisation of Systems and Services (ITMS 26240220072), Knowledge discovery (ITMS 26240220063) co-funded by the ERDF.

References

1. Tatara, N., Arsand, E., Nilsen, H., Hartvigsen, G.: A review of mobile terminal-based applications for self-management of patients with diabetes. In: Intl. Conference on eHealth, Telemedicine, and Social Medicine, Mexico, pp. 166–175 (2009)
2. NHS Choices UK's health website, <http://www.nhs.uk>
3. Friedman, C.: Semantic Text Parsing for Patient Records. Published in Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Springer (2005)
4. Friedman, C., Johnson, S.B.: Natural Language and Text in Processing in Biomedicine. In: Biomedical Informatics: Computer Applications in Health Care and Medicine. Springer (2005)
5. Aronsky, D., Fiszman, M., Chapman, W.W., et al.: Combining decision support methodologies to diagnose pneumonia. In: Proc. AMIA Symp., pp. 12–16 (2001)
6. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.: What can natural language processing do for clinical decision support? *J. Biomed. Inform.* 42, 760–772 (2009)
7. Dupuis, E.A., White, H.F., Newman, D., et al.: Tracking abnormal cervical cancer screening: evaluation of an EMR-based intervention. *J. Gen. Intern. Med.* 25, 575–580 (2010)
8. Waghlikar, K.B., MacLaughlin, K.L., Henry, M.R., Greenes, R.A., Hankey, R.A., Liu, H., Chaudhry, R.: Clinical decision support with automated text processing for cervical cancer screening. *J. Am. Med. Inform. Assoc.* (Published Online First April 29, 2012)
9. Chen, H., Fuller, S.S., Friedman, C., Hersh, W.: Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Integrated Series in Information Systems, vol. 8. Springer (2005)
10. McDonal, D.M., Su, H., Xu, J., Tseng, C.-J., Chen, H.: Gene Pathway Text Mining and Visualization published in Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Springer (2005)
11. Buchholz, S.N.: Memory-Based Grammatical Relation Finding. *Computer Science* 2, 17 (2002)
12. Krajčí, S., Novotný, R.: Lemmatization of Slovak words by a tool Morphonary, Tools for Acquisition, Organisation and Presenting of Information and Knowledge (2). In: Proceedings in Informatics and Information Technologies, Vydavateľstvo STU, pp. 115–118 (2007) ISBN 978-80-227-2716-7