Integrating Patient-Related Entities Using Hospital Information System Data and Automatic Analysis of Free Text

Svetla Boytcheva, Galia Angelova, Zhivko Angelov, Dimitar Tcharaktchiev, Hristo Dimitrov

Integrating Patient-Related Entities using Hospital Information System Data and Automatic Analysis of Free Text

Svetla Boytcheva[1], Galia Angelova[1], Zhivko Angelov[1], Dimitar Tcharaktchiev[2], Hristo Dimitrov[2]

[1]Institute of Information and Communication Technologies (IICT), Bulgarian Academy of Sciences, Sofia, Bulgaria
[2]University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev" (USHATE), Medical University Sofia, Bulgaria

Abstract: The article presents research in secondary use of information about medical entities that are automatically extracted from the free text of hospital patient records. To capture patient diagnoses, drugs, lab data and status, four extractors that analyse Bulgarian medical texts have been developed. An integrated repository, which comprises the extracted entities and relevant records of the hospital information system, has been constructed. The repository is further applied in experiments for discovery of adverse drug events. This paper presents the extractors and the strategy of assigning time anchors to the entities that are identified in the patient record texts. Evaluation results are summarised as well as application scenarios which make use of the extracting tools and the acquired integrated repository.

Keywords: automatic information extraction; secondary use of patient records; temporal aspects of data integration

1 Introduction

Electronic Health Record (EHRs) are viewed as the basic source of patient-related data, keeping all important medical information about each patient and (in a longer run) providing access to the complete patient history. The idea to re-use the EHR content beyond the direct health care delivery is relatively recent (published in 2007, its implementation is still in its infantry according to [1]). EHR data can facilitate the clinical research and reduce substantially the cost of clinical trials as they provide an enormously large resource for statistical observations, comparative studies, quality evaluation, monitoring the effectiveness of public health services and so on. By default the Information Technologies (IT) are the only means to cope with the large data volumes, moreover EHRs are to be supported within IT environments which provide secure, confidential, and private data access. Therefore, it is important to design research prototypes where secondary EHR use demonstrates the potential of novel, IT-enabled developments for improving the clinical practice.

This ambitious and challenging objective, however, faces the incompleteness, fragmentariness, inconsistency and vagueness of expressions in the established medical practices, which rely on domain and/or implicit knowledge to support the information interpretation. Patient-related data are recorded in various formats, encoded by numerous medical nomenclatures and classifications, with heterogeneous terminologies, specific national traditions to shape the patient record texts and so on. Much EHR information is presented as free text so the Natural Language Processing (NLP) is often viewed as an embedded technology that helps to extract structured knowledge chunks from the EHR texts. Current tasks related to secondary EHR use seem to be mainly focused on the extraction activities; there are fewer integration attempts which aim at the exploitation of the accumulated information.

This paper presents an experimental integration of patient-related clinical data in order to construct a repository for identification of Adverse Drug Events (ADE) in the PSIP project (Patient Safety through Intelligent Procedures in Medication) [2]. Components for automatic extraction of entities from free text have been developed since much information is documented in Bulgarian hospitals as unstructured text. The article discusses specific solutions regarding time anchoring and harmonisation of data units while the automatically extracted entities are integrated with the Hospital Information System (HIS) records to form a unified patient case.

The article is structured as follows. Section 2 overviews related research. Section 3 presents the background of our experiment. Section 4 considers our approach to integration of all recognised entities, which are available as structured HIS values or are extracted by the text analysers. Event sequencing is important and we present our empirical strategy for assignment of temporal markers to all findings. Section 5 considers the evaluation results and discusses feasible application scenarios of the integrated repository given that it inevitably contains some inconsistencies and certain percentage of erroneous assignments. Section 6 contains the conclusion.

2 RELATED WORK

Data quality issues, related to secondary EHR use, are discussed in [3]. The authors consider three categories of data quality: *(i) incompleteness* – missing information; *(ii) inconsistency* – information mismatch between various or within the same EHR data source; and *(iii) inaccuracy* – non-specific, non-standards-based, inexact, incorrect, or imprecise information. The article [3] reports about inconsistencies which are common to many data collections, e.g. 48% of the patients did not have corresponding diagnoses or disease documentation in the pathology reports. The suggestion is to develop software tools for automatic data validation and flexible data presentation in order to support information integrity. This recent article encourages us to continue our experiments despite the negative data quality findings in our training corpus; we apply the data quality categories introduced in [3].

Rule-based automatic approaches for data extraction and integration ensure the state-of-the-art achievements in the construction of large scale resources (over 300 millions rows of data from three institutes are currently included in the Biomedical Translational Research Information System [4]). The mapping rules are created manually based on human analysis, using a large dictionary of medical terminology.

Predefined queries are developed in the system to respond information needs. This article discusses various kinds of (potential) application of such an integrated resource which however is difficult to obtain as it requires long years of data collection.

Modelling of timelines is considered in [5] which overviews the six state-of-the-art systems related to visualisation of temporal information in EHRs. Most of these systems operate on readily available lists of type- and time-tagged events. One of the systems identifies pre-defined classes of entities (e.g. diseases, investigations, problems, drugs, etc.) and semantic relationships between them (e.g. investigation indicates problem) in natural language texts [6]. Searching useful information with self-service visual query tool is implemented in repositories containing preliminary indexed full-text documents, e.g. the system STRIDE works on a clinical data warehouse containing information about over 1,3 million patients [7].

As our extractors perform free text analysis on the raw texts in the USHATE HIS, we briefly overview the recent achievements in the area.

We deal with the automatic assignment of ICD-10 diagnoses to free text phrases in Bulgarian language (ICD-10 is the International Classification of Diseases, version 10 [8]). A recent review [9] summarises achievements in the automatic coding of English medical texts. The authors note that software tools for automating coding are "available but currently not widely used, most likely because the systems are still in development and their performance in production is unproven". Various evaluation metrics are summarised and the best result reported is 98% coding accuracy. We note that systems coding English medical texts are developed since more than 20 years; results for other natural languages are less precise; e.g. for French the agreement between the automatic procedure for assignment of ICD 10 codes and the EHR content is 21%, which is partly due to the fact that the diagnoses encoded in the hospital information systems often reflect financial considerations [10].

Another important extraction task is the automatic recognition of drugs and dosages, which occur in the patient record texts. State-of-the-art results reported for English are: sensitivity/recall for drug names 88,5% and for dosage 90,8%; precision for drug names 91,2% and for dosage 96,6% [11]. A measure that combines the sensitivity (recall) and the precision is their harmonic mean f-score; another highly successful extraction system is MedEx [12] which extracts drug names with f-score 93,2%, and achieves f-scores 94,5% for dosage, 93,9% for route and 96% for frequency. The French Multi-Terminology Indexer *F-MTI* [10] achieves very good results in drug extraction as well. Codes from the ATC (Anatomical Therapeutic Chemical) classification [13] are automatically assigned to the extracted drug names. The extraction of ATC codes from the free text of French discharge letters is performed with f-score 88% when compared to the manual extraction; however, compared to the Hospital Pharmacy content, the f-score is 49%. These figures are the baseline for assessment of our results in the automatic recognition of drugs in Bulgarian patient record texts.

The performance of information extraction from clinical texts gradually improves and exceeds 90% accuracy [14]. In our experiments we have also developed different modules which focus on various text entities: patient status and values of clinical tests and lab data. These modules were implemented within a period of two years with progressive collection of the corresponding lexical and grammatical resources, leading to progressive improvement. The extractors were trained via specific procedures and can be applied as separate text analysis components. Our results are comparable to the state-of-the-art achievements in the area, see section 5.

3 RESEARCH CONTEXT

Our experiments in secondary use of EHR content are performed on training and testing corpora which contain anonymised hospital Patient Records (PRs).

3.1 USHATE - Clinical Settings

USHATE belongs to the oldest and largest Medical University in Bulgaria and is specialised mostly for treatment of endocrine and metabolic disease. Usually patients with such type of diseases have many complications and accompanying illnesses. In this way many patients arrive to the hospital with drugs prescribed elsewhere (in some other hospital, for instance, or at the ambulatory care, or from their GPs). Statistical observations show that the average number of drugs, discussed in a patient discharge letter in USHATE, is 5,4 drugs per hospital record. However, according to the Computerised Physician Order Entry (CPOE), there are fewer medications given to the patients: 1,9 per hospital record. The drugs for accompanying and chronic diseases, which are not prescribed via the Hospital Pharmacy, are entered in the discharge letters as free text. Thus, much information about the treatment is presented as unstructured descriptions. Similar comments can be made concerning the clinical examinations and lab tests: often the patients bring their test results on paper; the practice is not to repeat recent tests. Whenever the clinical test is made in USHATE, the lab data are stored automatically in the Laboratory Information System (LIS) which is a part of the HIS. However the values of clinical tests, made outside USHATE, are re-typed to the discharge letter for all examinations that are relevant to the present hospitalisation. Another particularity is due to the fact that reimbursements by the Bulgarian National Health Fund are based on clinical pathways; often the USHATE experts diagnose formally the principal disease which is sufficient to associate the patient to the desired pathway. But the hospitalised patients usually have specific, complex history cases. Statistical observations show that the average number of diagnoses per patient is 4,32 in contrast of only 1 diagnose recorded in HIS. Thus accompanying illnesses and complications, which are not formally encoded by the USHATE HIS, might be enumerated in the text paragraphs of the discharge letters. In this way the integrated picture of the patient diseases, history, status, lab data and treatment is presented only in the discharge letters of USHATE patients.

Therefore the secondary use of the hospital PRs in USHATE requires: *(i)* obligatory *text analysis* – to identify in the text the drugs, diagnoses, values of clinical tests, lab data as well as patient status attributes; *(ii)* strategies for *integration* of the various information fragments and *(iii)* maintenance of *incomplete information* – including timing of various events which are not precisely dated in the free text.

3.2 Specific Text Features of Bulgarian Discharge Letters

Discharge letters in all Bulgarian hospitals have mandatory structure, which is published in the Official State Gazette within the legal Agreement between the Bulgarian Medical Association and the National Health Insurance Fund [15]. The letters contain: *(i)* personal data; *(ii)* diagnoses; *(iii)* anamnesis; *(iv)* patient status; *(v)* lab data; *(vi)* medical examiners comments; *(vii)* discussion; *(viii)* treatment; and *(ix)* recommendations. This structure provides appropriate contextualization of the

extracted entities although it is often neglected because the authoring experts merge sections, omit empty sections, skip subsections, enter incomplete information and so on. Fig. 1 shows the percentage of existing sections in a training corpus of 1,300 anonymised PRs. After our studies in literature dealing with biomedical NLP for other languages, we note that the relatively established format of Bulgarian discharge letters is a significant advantage in the development of automatic analysis for Bulgarian.

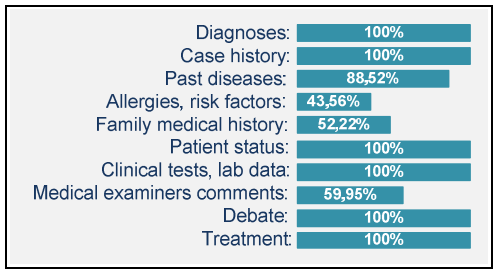

Figure 1. Availability of discharge letter zones

Finally we briefly present some specific particularities of the Bulgarian medical language. In the hospital PRs, medical terminology is recorded in both Bulgarian and/or Latin language. There is no preferred language for the terminology so the two forms are used like synonyms. Sometimes Latin terms are written by Cyrillic letters especially when the medical expert prefers to avoid keyboard switching. In general the mixture of Latin and Bulgarian terms is traditionally established and commonly accepted, including in official documents. In this way the automatic identification of a term in the hospital PR texts is a tricky task which requires more than a simple string match. Fig. 2 shows some original excerpts of PR diagnoses: Latin names of diseases are transliterated by Cyrillic letters but alternatively might be given in Latin as well. The measurement units of clinical test are often entered with Latin symbols.

```
... Диабетес мелитус – типус 1. Ретинопатиа диабетика пролиферанс.
Статус пост ALC. Полиневропатиа диабетика. Хипертония артериалис гр.
I. ...

... Консултация с офталмолог: VOD= 0,8; VOS= 0,6-0,7; двуочно 0,9-1,0
със собствена корекция. Angiosclerosis vas. retinae hypertonica.
Retinopathia diabetica simplex. макули без рефлекс...
```

Figure 2. Excerpts of hospital PRs in USHATE

4 Integration of Entities Encountered in Hospital PRs

For our PSIP experiment, all entities automatically extracted from the hospital PRs should be integrated as complementary data to the information, contained in the HIS (including the LIS, the CPOE and the Hospital Pharmacy). Therefore we need to design the time framework which unifies and harmonises all events encountered in every particular discharge letter.

The dates (beginning and end of the hospital stay) are set in the HIS of USHATE for each patient. The admission day is referred to as 'Day 0'. All drug prescriptions and lab data, stored in the HIS, are marked by respective dates and hours within the

hospital stay interval so the time markers of these medication and examination events are delivered by the HIS. Fig. 3 illustrates the association of time anchors to the entities, extracted from the PR text by the four extractors in our experiment:

(i) *Diagnoses extraction module* – automatically extracts diagnoses from the PR zone "Diagnoses", associates the corresponding ICD-10 codes and stores them in the PSIP repository;

(ii) *Drug events extraction module* – automatically extracts drug names, recognised as "current treatment of accompanying diseases", and their dosage. The module automatically associates ATC codes to these drugs. In case they are not prescribed by the Hospital Pharmacy, the extractor relates them to Day 0 (and for chronic diseases like e.g. hypertonia, to each day spent in the hospital day). Then these drugs are stored for the current patient in the PSIP repository;

(iii) *Clinical data extraction module* - automatically analyses the clinical examinations and tests, made outside USHATE (e.g. hormonal tests), anchors the to Day 0 and stores them in the PSIP repository;

(iv) *Status extraction module* – automatically extracts the patient status' attributes, which by default reflect the patient characteristics at Day 0.

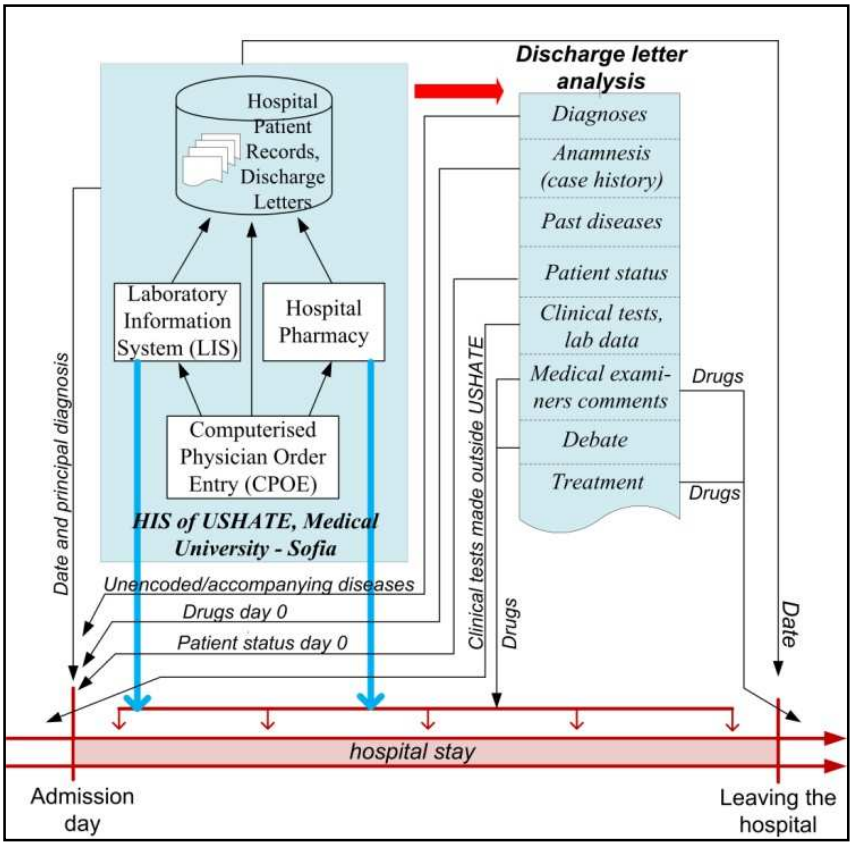

Figure 3. Integration of data and events stored in the Hospital Information System of USHATE and in the free texts of discharge letters

The clinical treatment usually starts at Day 1, so we have made experiments to discover automatically the drugs taken at Day 0. Only the text paragraphs in the *Anamnesis* were considered because this is the PR zone where drugs are listed, with comments that some drugs are taken at the hospitalisation moment. Dosage is often specified explicitly as definition of the admission medication status. The experiment was carried out with the corpus of 6,200 PRs and all 1,537 drugs in our lexicons. In general there are many drugs presented in the *Case history* (30,190 drug name occurrences in 6,194 *Anamneses*). Some 4,088 occurrences (13,5%) were automatically recognised as "current medication" in 3,191 PRs, which means that 51% of the PRs contain explicit statements concerning the "current medication" in the *Anamnesis* text. More than one drug can be specified as a "current" one.

Most often the text expression, signalling the treatment at Day 0, is the phrase "at the admission" (*при постъпването*). This phrase occurs with slight variations in 2,122 PRs (34%). On average 25% of all drugs in the *Case history* are recognised as "Day 0 medication". There are PRs, however, for which up to 45% of the medication events listed in the *Case history* are "current". A sample PR excerpt is: "therapy at the admission novorapid 14+14+14E, insulatard 24E at 22h" (*терапия при постъпването новорапид 14+14+14E, инсулатард 24E в 22ч*)".

The second preferred text expression which signals "Day 0 medication" is "*at the moment*" (*в момента*...). It occurs in 908 PRs from the test set of 6,200 PRs; in 703 PRs (11%) the phrase refers to explicitly specified drugs in the local context. These 703 cases were encountered after manual inspection of automatically prepared summarising sheets for occurrences of drug names in the *Anamneses*. On average, 14,5% of all drugs in the 703 PRs *Case history* are recognised as "Day 0 medication". Sample usages of this phrase are: "at the moment takes dostinex 2x2 t weekly" (*в момента на достинекс 2x2 т седмично*) and "at the moment treated with physiotens 0,4mg, lorista 1t, isoptin 2x80mg with satistactory effect".

The above-listed drugs are considered as Day 0 only after careful training of the extracting components and evaluation of the erroneous recognitions. Please note that the expressions "*at the admission*" and "*at the moment*" can be used in other phrases as well, like "*therapy at the admission: none*", "*at the moment without complains*" and "*aged X, at the moment 93 kg*".

As an example of integrated entities, we present the Patient Case with ID 26137: *(i)* the principle diagnose in the USHATE HIS is E668 '*Other obesity*', and the extractor discovers in addition E898 *'Postprocedural adrenocortical (-medullary) hypofunction'* and E289 *'Other ovarian dysfunction'*; *(ii)* there is only one entry found in the Hospital Pharmacy (*Metamizole*) and the drug extractor delivered eight records for taking *Metfodiab* for eight days; *(iii)* there are 22 lab data entries in the USHATE HIS but the mining component adds seven more values at Day 0 mostly for hormones.

4.1 Automatic Extraction of ICD-10 Codes to Diagnoses

The PRs in our corpus often contain diagnoses expressed as free text (despite the fact that the present USHATE HIS offers menu choice for ICD-10 diagnoses). For the corpus of 6,200 PRs, some 495 different ICD-10 codes were identified in the PR zones *Diagnoses*. Fortunately the zones themselves are presented as major discharge letter paragraphs and can be recognised automatically with almost 100% accuracy. However, in the PR text we find terms, phrases and paraphrases which might differ significantly from the ICD disease labels. The number of diagnoses per patient varies from 1 to

about 30; most PRs contain from one to seven diagnoses. The nomenclature ICD-10 with Bulgarian terms is used as an external resource and lexicon for this module.

The main obstacle for the automatic assignment of ICD codes is that the correspondence between the PR diagnoses and ICD-10 disease names is "many to many". There are generally formulated phrases in the PR text which correspond to several ICD-10 diagnoses (like e.g. *hypothyroidism*). Some phrases can be matched exactly to one ICD-10 disease (for instance, *diabetic polyneuropathy*). Major difficulties for the linguistic analysis are caused by transliterations and paraphrases: *(i)* there are Latin names of the illness, transliterated by Cyrillic letters, which differs from the Bulgarian labels included in the ICD-10, for instance, *'Диабетес мелитус'* (*Diabetes Mellitus*) in the PR text vs *'Захарен диабет'* in ICD; *(ii)* the PR texts contain syntactical paraphrases of disease names, often with mixture of Bulgarian-Latin writing, and/or join of various diseases and symptoms in conjunctive nominal phrase, for instance *'полиневропатия ет нефропатия диабетика в стадий на микроалбуминурия'* (*polyneuropathy et neuropathy diabetica in the stage of microalbuminuria*); *(iii)* the PR texts contain sophisticated syntactic constructions, e.g. splitting the components of complex ICD-10 terms and presenting them into various levels of a joint syntactic structures.

For solving this task a machine learning algorithm was especially designed and implemented. A smaller training corpus of 600 PRs was used for manual association of corresponding ICD-10 codes to diagnoses as they are presented in the text. Further the algorithm is automatically trained on 1,300 PRs. More details about the automatic coding of ICD-10 diagnoses in Bulgarian are given in [16].

4.2 Automatic Extraction of Drugs

The list of drugs in the USHATE's Hospital Pharmacy is supported with Bulgarian drug names even for drugs produced abroad (in this case the foreign words are transliterated by Cyrillic letters). However, the official list of registered drugs, published by the Bulgarian Drug Agency [17], contains the ATC codes and the drug names in Latin alphabet even for drugs produced in Bulgaria. It is worth mentioning that all the *Application instructions* in the Bulgarian Drug Agency site [17] are written in Bulgarian and the drug names are given there by Cyrillic letters. Note that the ATC classification is not available for drug names in Bulgarian language; therefore we have selected about 2000 drug names (covering the drugs relevant for the USHATE patients in the PSIP corpus) and have (semi-)manually assigned ATC codes to drug names in Bulgarian. In the process of resource compilation for the corpus of 6,200 discharge letters, it became clear that the USHATE patients take 355 drugs during the hospitalisation period, which are not prescribed via the Hospital Pharmacy. The drug extractor is focused on identification of these drugs when they are taken during the hospitalisation period. The information extraction task is accomplished by a rule-based algorithm that uses over 50 regular expressions for drug events recognition.

Recognising drug names is based on string matching which is difficult due to many reasons. Drugs have various names that might be referred to in the PR texts: e.g. brand and generic names. There are variants in writing names, especially for names consisting of several strings. Actually multi-word drug names might occur in the PR text as a single wordform because the other name parts are omitted. Additionally, drug names in the PR might be written with Cyrillic letters, for Bulgarian names and transliterated Latin names, and with Latin alphabet. In order to capture all names

during the text processing phase, we need a comprehensive vocabulary of drug names in both languages and both alphabets **(**as well as the ATC codes).

Due to the highly inflectional Bulgarian language, some drug names might appear in the PR texts with various wordforms. This lexical variety, which prevents the exact match of drug names from the PR texts to the lexicon items, occurs mostly for plural and singular forms. In 1,300 PRs, some 43 grammatical forms of 23 drug names were automatically found by our morphological analysers. Another obstacle in string matching might be due to the typos. In principle spelling errors prevent the correct recognition of all text entities in the PR texts (and need to be tackled by automatic correctors). For 1,300 PRs, 4,042 drug name occurrences of 239 drugs were automatically identified. Some 100 PRs were manually studied for typos. The erroneous occurrences are 59 (1,5%) and the errors appeared in the names of 21 drugs.

The lexicons of our drug extracting component contain 1,182 drug names, which are prescribed via the USHATE Hospital Pharmacy, and another 355 drugs that occur in the USHATE hospital PRs but are not prescribed via the Hospital Pharmacy (the latter are taken by the patients to cure additional/chronic illnesses). Tokens which are part of drugs names occur in the whole PR text; in fact drug names participate even in the zone *Diagnoses* (e.g. 'deficiency of Vitamin D'). In this way our procedure for automatic recognition of drug names finds words, signaling potential drug treatments, everywhere in the PRs. More details about the drug extraction component can be found in [18]. The contextualisation of medication events (i.e. to recognise the drugs admitted during the hospitalisation period) is further discussed in [19]. In section 5 we present new evaluation results concerning the extraction of “current” medication events from the unstructured texts of the PR *Anamnesis.*

4.3 Automatic Extraction of Values of Clinical Tests and Lab Data

Fig. 1 shows that the lab data are presented in a specific PR zone which is practically always available and can be automatically identified with almost 100% accuracy despite the variety of section titles and subtitles. The values are listed without predetermined order, using measurement units and their abbreviations both in Bulgarian and Latin. These measurement units are compliant to the LOINC (Logical Observation Identifiers Names and Codes) [20] classification and often enable the recognition of the corresponding indicator which might be referred to without explicit and standardised indicator name. The lab data extractor identifies the tested attribute and its value. The units and reference intervals are desirable features to recognise, and the time, condition and explanation of further details are optional features. The extraction algorithm is based on rules and pattern matching; the rules are acquired after manual training on 1,300 PRs and recorder in different versions to cope with various delimiters and blank spaces, which might occur in the text [16].

5 Evaluation Results and Discussion

The automatic extractors were run on 6,200 anonymised PRs. The accuracy for the automatic extraction of diagnoses, drugs, and clinical tests data is presented in Table 1. These events were integrated with the HIS data to constitute the PSIP repository [16]. Recently the repository was used for discovery of USHATE-specific ADEs.

	Extracted entities from the PRs text	Precision	Recall	F-Measure
Diagnoses	26 826	97.30%	74.69%	84.50%
Drug names	160 892	97.28%	99.59%	98.42%
Laboratory Test Results	114 441	98.20%	99.99%	99.04%

Table 1. Number of extracted items from 6,200 PRs

The medication events occurring during the hospitalisation are recognised with f-score 90.17% for 355 drugs [19]. The over-generation is 6%, i.e. some drugs are wrongly classified as "admitted during the hospital stay". These erroneous decisions are made for phrases like "… *to continue the treatment with drug X* …" in the *Debate* section, which communicate incomplete information and are ambiguous for human being as well. In all cases of overlapping descriptions the HIS data are preferred as more exact and reliable. In general the automatically extracted entities have mostly statistical meaning in the procedure of data analysis and ADE discovery for USHATE.

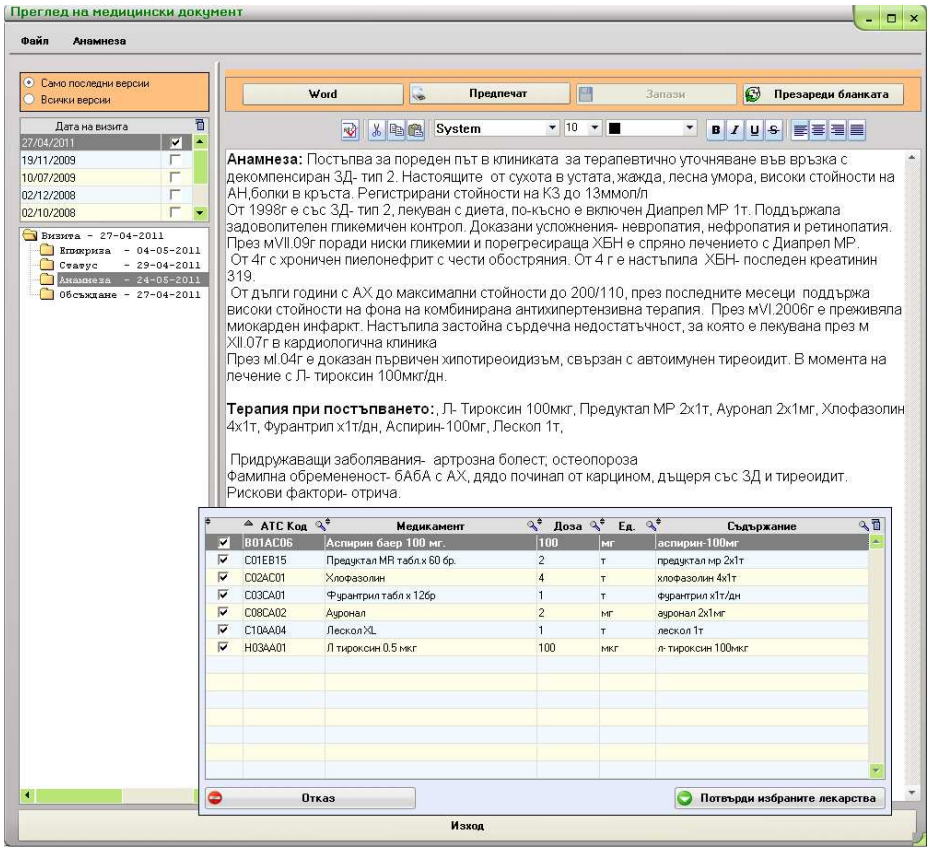

Figure 4. Integration of the drug extractor into on-line validation interface in USHATE

Recently an experimental validation of the PSIP Scorecards [21, 22] in USHATE has been accomplished. Actually the extractors, developed for the Bulgarian PR texts, provided interoperability between the USHATE PRs and the PSIP data formats: once structured information is extracted from the free texts, it can be recorded in various databases using ATC and ICD codes. The validating doctors were quite positive about the experimental integration of the drug extractor as an on-line analyser in the HIS (see Fig. 4). It delivers automatically the drugs taken at Day 0 whenever the *Anamnesis* is recorder in the HIS, which provides structured data in a convenient format that can be further used for prescriptions. Validating the PSIP approach, some situations at risk were found in the experimental USHATE repository (hypo- and hyperkalemia, hypo- and hyperglycemia, leucopenia, renal failure). The integrated repository of patient data, prepared using the technologies presented in this article, is an example of resource which explicates the potential of secondary EHR use.

6 Conclusion

This article presents a research effort in automatic extraction of structured information from hospital PRs, performed in order to integrate a repository for experimental discovery of ADEs. We have described our empirical strategy to assign time anchors to the entities encountered in the PR free texts. The integrated repository for USHATE is relatively small but relatively sophisticated as it comprises the HIS data as well as the results of four automatic extractors.

The information extraction approach is to identify entities of interest and to implement software tools which perform partial analysis of the text fragments that contain words of interest. The remaining part of the PR texts is disregarded. Our experience shows that via a rapid development process, one can achieve good performance in several automatic extraction tasks within 2-3 years. To some extent the extraction accuracy reported here is implied by the established structure of the discharge letters in Bulgarian hospitals. The negative results (including over-generation) are an inevitable aspect of the NLP performance but they are partly due to the inconsistency, incompleteness and fragmentariness of the medical documentation per se; these shortcomings become obvious in the computer age when ambitious goals like secondary use of EHR data are set. The false positive indications might be dangerous for further use but in our case the small percentage of false positive entities is statistically insignificant and practically negligible (we note that human recognition of medical entities in clinical narratives and data preparation might also include some erroneous choices).

We also note the stable medical tradition to type in textual descriptions even when the HIS stores the prescriptions. Medical experts document carefully information about the therapy and its changes during the hospital stay. These practices make the NLP technology a valuable component in the secondary use of EHR data.

7 Acknowledgment

The research tasks leading to these results have received funding from the EC's FP7 ICT under grant agreement n° 216130 PSIP (Patient Safety through Intelligent Procedures in Medication) as well as from the Bulgarian National Science Fund under

grant agreement n° DO 02-292 EVTIMA (Efficient Search of Conceptual Patters with Application in Medical Informatics).

8 References

[1] Prokosch H. and T. Ganslandt. *Perspectives for medical informatics. Reusing the electronic medical record for clinical research.* Methods Inf. Med. 2009; 48(1), pp. 38-44.

[2] PSIP (Patient Safety through Intelligent Procedures in Medication), www.psip-project.eu.

[3] Botsis, T., G. Hartvigsen, F. Chen, and C. Weng. *Secondary Use of EHR: Data Quality Issues and Informatics Opportunities*. AMIA Summits Transl. Sci. Proc. 2010; pp. 1–5.

[4] Cimino J. and E. Ayres. *The clinical research data repository of the US National Institutes of Health*. Stud. Health Technol. Inform. 2010; 160(Pt 2): pp. 1299-1303.

[5] Roque, F., L. Slaughter, and A. Tkatchenko. *A Comparison of Several Key Information Visualisation Systems for Secondary Use of EHR Content*. In Proc. NAACL HLT 2nd Louhi Workshop on Text and Data Mining of Health Documents, June 2010, pp. 76–83.

[6] Hallett, C. *Multi-modal presentation of medical histories*, IUI '08: Proc. 13th Int. Conf. on Intelligent User Interfaces, ACM New York, 2008, pp. 80-89.

[7] Lowe H., T. Ferris, P. Hernandez, and S. Weber. *STRIDE - An integrated standards-based translational research informatics platform*. AMIA Annual Symp. Proc. 2009, pp. 391-395.

[8] International Classification of Diseases (ICD), WHO, http://www.who.int/classifications/icd/en/

[9] Stanfill, M., M. Williams, S. Fenton, R. Jenders, and W. Hersh. *A systematic literature review of automated clinical coding and classification systems*. JAMIA 2010 (17), pp. 646-651.

[10] Merlin B., E. Chazard, S. Pereira, E. Serrot, S. Sakji, R. Beuscart, and S. Darmoni. *Can F-MTI semantic-mined drug codes be used for Adverse Drug Events detection when no CPOE is available?* Stud. Health Technol. Inform. 2010; 160(Pt 1): pp. 1025-1029.

[11] Halgrim, S., F. Xia, I. Solti, E. Cadag, and Ö. Uzuner. *Extracting medication information from discharge summaries*, In Louhi '10 Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, 2010, 61-67.

[12] Xu.H., S. P Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny. *MedEx: a medication information extraction system for clinical narratives*. JAMIA 2010, 17, pp. 19-24.

[13] ATC drugs classification, http://www.whocc.no/atc_ddd_index/

[14] Meystre, S., G. Savova, K. Kipper-Schuler, and J. F. Hurdle. *Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research*, IMIA Yearbook of Medical Informatics 2008, pp. 138-154.

[15] National Framework Contract between the National Health Insurance Fund, the Bulgarian Medical Association and the Bulgarian Dental Association, *Official State Gazette* №106/30.12.2005, updates №68/22.08.2006 and №101/15.12.2006, Sofia, Bulgaria, http://dv.parliament.bg/.

[16] Tcharaktchiev, D., G. Angelova, S. Boytcheva, Z. Angelov, and S. Zacharieva. *Completion of Structured Patient Descriptions by Semantic Mining*. Stud. Health Technol. Inform. 2011, 166: 260-269.

[17] Bulgarian Drug Agency, http://www.bda.bg/index.php?lang=en.

[18] Boytcheva, S. *Shallow Medication Extraction from Hospital Patient Records*. Stud. Health Technol. Inform. 2011, 166: pp. 260-269: pp. 119-128.

[19] Boytcheva, S., D. Tcharaktchiev and G. Angelova. *Contenxtualisation in Automatic Extraction of Drugs from Hospital Patient Records*. To appear in the *Proc. of MIE-2011, the 23th Int. Conf. of the European Federation for Medical Informatics*, Norway, 28-31 August 2011, published by IOS Press.

[20] Logical Observation Identifiers Names and Codes (LOINC®), http://loinc.org/

[21] Marcilly, R., E. Chazard, M.-C. Beuscart-Zéphir, W. Hackl, A. Baceanu, A. Kushniruk, and E. Borycki. *Design of Adverse Drug Events-Scorecards*. Proc. Int.l Conf. .Information Technology and Communications in Health (ITCH), 2011, Victoria, CA.

[22] Koutkias, V., V. Kilintzis, G. Stalidis, K. Lazou, C. Collyda, E. Chazard, P. McNair, R. Beuscart, the PSIP Consortium and N. Maglaveras. *Constructing Clinical Decision Support Systems for Adverse Drug Event Prevention: A Knowledge-based Approach*. AMIA Annu Symp Proc. 2010, 402–406.