Automatic extraction of ICD-10 Codes for Diagnoses and ATC Codes and Dosages for Drugs

Svetla BOYTCHEVA

Institute of Information and Communication Technologies (IICT), Bulgarian Academy of Sciences and State University of Library Studies and Information Technologies, Sofia, Bulgaria

svetla.boytcheva@gmail.com

Abstract

This paper presents approach for automatic extraction of diagnoses and medication information from discharge letters. The proposed algorithms are designed for processing free text document in Bulgarian language. Two algorithms were designed and implemented: an algorithm for mapping of International Classification of Diseases 10th revision (ICD-10) to diagnoses extracted from patient records (PRs) and an algorithm for automatic extraction of drug events and association to them of codes from Anatomical Therapeutic Chemical (ATC) Classification System. The system presented in here was developed and applied in the PSIP project for the preparation of an experimental repository for PSIP validation from University Specialized Hospital for Active Treatment of Endocrinology "Acad. I. Penchev" (USHATE) at Medical University – Sofia.

Keywords

Medical Information Extraction, Semantic mining

1. Introduction

We deal with hospital Patient records (PRs) which are anonymized by the hospital information system of the University Specialized Hospital for Active Treatment of Endocrinology "Acad. I. Penchev" (USHATE) at Medical University – Sofia.

The system presented in here was developed and applied in the PSIP project for the preparation of an experimental USHATE's repository for PSIP validation [1].

PRs 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 [2].

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.

2. Drug Events Information Extraction

We have developed automatic procedures [3] for analysis of free texts in hospital patient records in order to extract information about: drug names; dosages; modes; frequency and treatment duration and to assign the corresponding ATC code [4] to each medication event.

Actually the system enables extraction of drug-related information about drugs which are mentioned in the PR texts as accompanying medications but are not prescribed by the Hospital Pharmacy.

The list of registered drugs in Bulgaria is provided by the Bulgarian Drug Agency [5] and it contains about 4,000 drug names and their ATC codes. Currently the Hospital Pharmacy operates with 1,537 medications because USHATE is specialized mostly for treatment of diabetic patients.

Our aim in the PSIP project is to extract information about drugs, taken by the patient, which are not prescribed via the Hospital Pharmacy.

There are two versions of the implemented algorithms for information extraction of drug events from discharge letters in Bulgarian language:

- Retrospective – This version was used in the initial preprocessing stage, where major task was

information collection from full text discharge letters included in USHATE archive (total 6200 PRs). The implemented system usually uses automatic model for analysis of all PRs in experimental repository (Fig. 10). For this task were extracted drug events from all sections of text personal records. In order to determine which drugs were used for treatment during hospitalization we maintained information from:

- ✓ *Anamnesis* – including history of disease and information for previous and current treatment;
- ✓ *Medical Examiners comments* – including information for treatment of accompanying diseases;
- ✓ *Debate* – including information for the provided treatment during hospitalization.

The extracted information concerning drug events from these three sections is processed as follows:

- ✓ If some drug is mentioned in the *Anamnesis* section, but not discussed further in *Debate* section, this means that treatment with it was stopped and we are not concerning this drug as "current treatment".
- ✓ If some drug is mentioned in the *Anamnesis* section and discussed further in *debate* section, this means this drug should be marked with "current treatment" marker. For such drugs we use last mentioned information about their dosage, because it can happened at hospital admission daily scheme to be changed. In such cases this new information is placed in *Debate* section. If the current treatment continues with the same scheme as admission we use information about dosage from *Anamnesis* section (if any)
- ✓ All drugs mentioned in *Medical Examiners comments* section are marked with "current treatment" marker, because they are recognized by the system with high confidence as present events (and not future prescriptions), except cases with 'stop' and 'replace' phrases.

- Prospective - This is the integrated version with Hospital information system at USHATE. There is also available manual mode of the system mainly used for testing and experiments (Fig. 1-9). For this task the algorithm provides runtime information extracted from *Anamnesis* section only, because the whole discharge letter is not available during the hospital admission. If anamnesis contains section "*Current treatment*", system extracts information available only in this section. In other cases we process temporal information concerning history of disease and treatment in *Anamnesis*. All text is separated by temporal events markers and segments between them are concerned as episodes describing diagnoses, treatment, symptoms and complain. The negation events are taken in consideration as well. Drugs in *Anamnesis*, only if they are listed under the headers *'medication at the moment of hospitalization'* or *'accompanying treatment'*. Some phrases like *'started treatment with'* can be also interpreted as hints for *'current medication'* but only if they are not followed by phrases including 'replaced by' which signal past events.

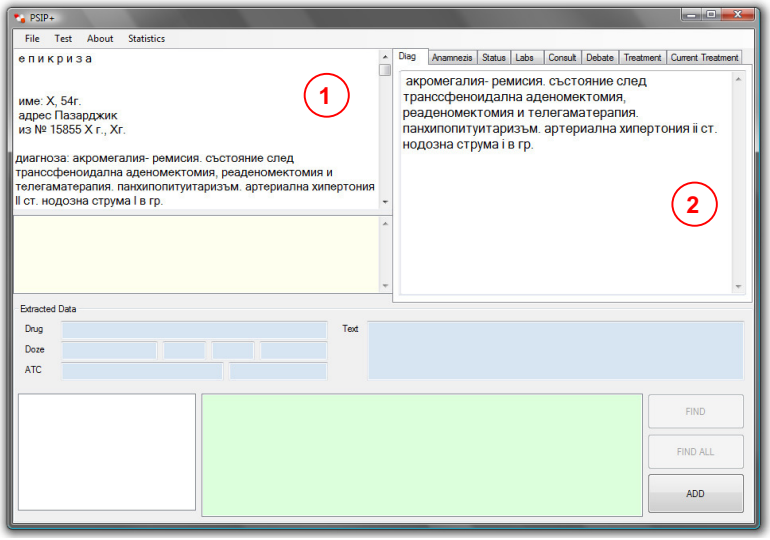


Fig. 1 Manual mode of the system

Allows processing of a single PR stored as text file. The text of PR is opened in section (1). After opening the text file, PR is automatically separated on sections and the text from each section is presented in separated tab (2): *(i)* personal data; *(ii)* diagnoses -Diag; *(iii)* anamnesis; *(iv)* patient status-Status; *(v)* lab data-Labs; *(vi)* medical examiners comments - Consult; *(vii)* discussion - Debate; *(viii)* treatment; and *(ix)* recommendations. The first section "personal data" is skipped, because we process anonymized PRs. The third section "anamnesis" is spited on two tabs – Current Treatment (if any) and Anamnesis. The information from the last two sections is merged and presented in the Treatment tab.

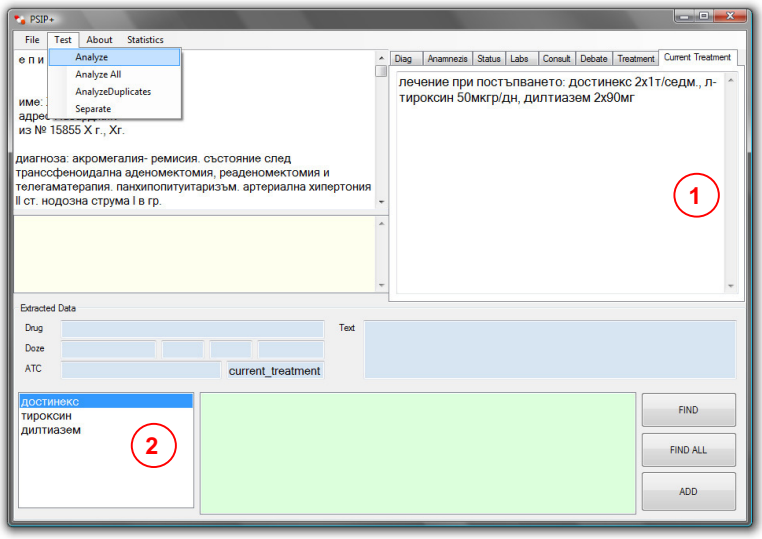


Fig. 2 Drug events manual analysis

PR section can be processed separately by the system. For the main task "identification of current treatment at hospital admission" we process information from "Current Treatment" section, if available or in other case from "Anamnesis". After choosing "Analyze" function from menu bar the selected section (1) is processed and automatically is generated list with recognized drug names within the text. In this example from "Current Treatment" section are recognized three drugs "достинекс" (Dostinex), „л-тироксин" (L-Thyroxin) and „дилтиазем" (Diltiazem). In this case all events are marked as present.

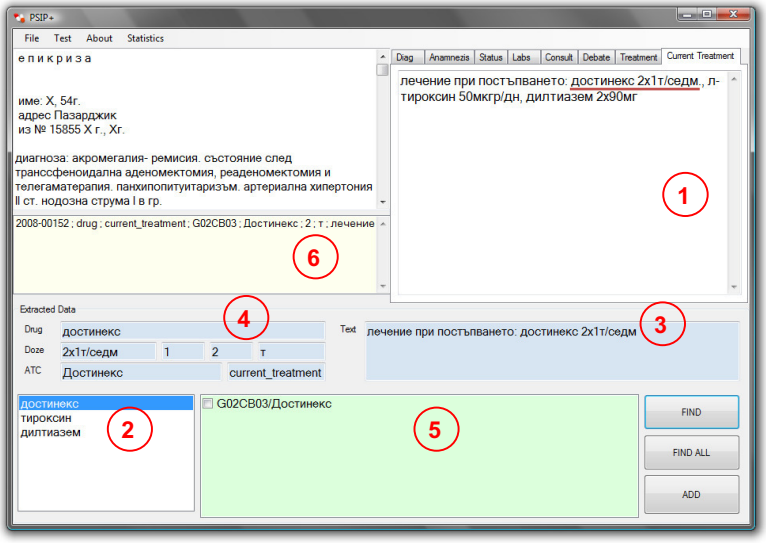


Fig. 3 Drug events Current Treatment

Drugs included in the generated list can be processed separately by using "Find" button and selecting drug name from the list or as bunch using "Find All" button.

In the current example is selected "достинекс" (Dostinex) from the list (2). The system identifies in the PR text (1) the scope of the corresponding drug event and presents it in section (3). Then using rules and regular expressions is identified information about drug name, dosages, modes, frequency and treatment duration (4). The system assigns the corresponding ATC code to this medication event (5) and stores information in CSV format (6).

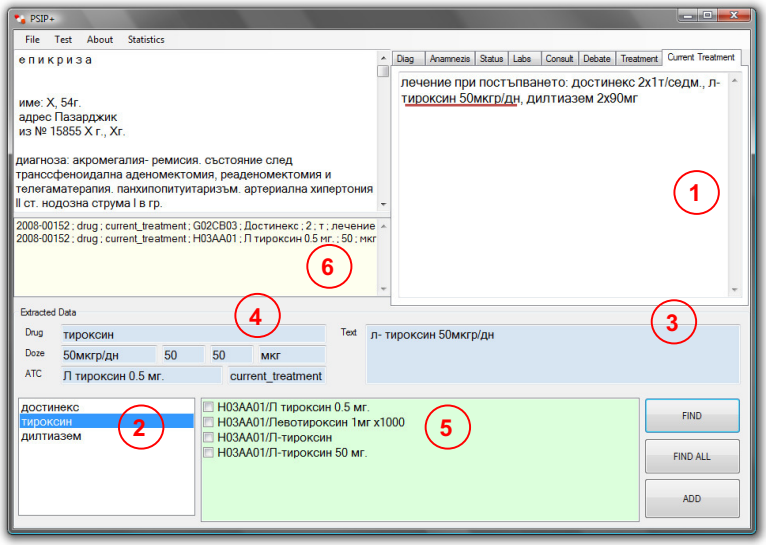


Fig. 4 Drug events disambiguation

In this example in check box list (5) are presented four possible drug packs for „л-тироксин" (L-Thyroxin) with 0.5 mg, 1 mg and 50 mg. Because in the PR text (1) the recognized dosage is 50 micrograms the closest dosage of 0.5 milligrams is selected from the list (5) and the ATC code of "л-тироксин 0.5 мг" (L-Thyroxin 0.5 mg) is associated to this drug event.

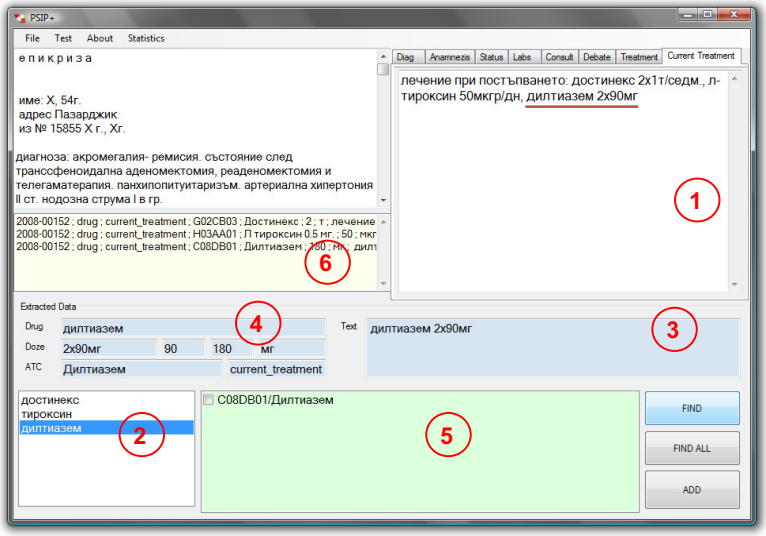


Fig. 5 Drug events daily dosage

In this example in check box list (5) is presented only one possible drug packs for „дилтиазем" (Diltiazem). Because there is no other options its ATC code is automatically is associated to this drug event. In the PR text (1) the recognized dosage is 2x90 mg per day and in section (4) is calculated the daily dosage 180 mg.

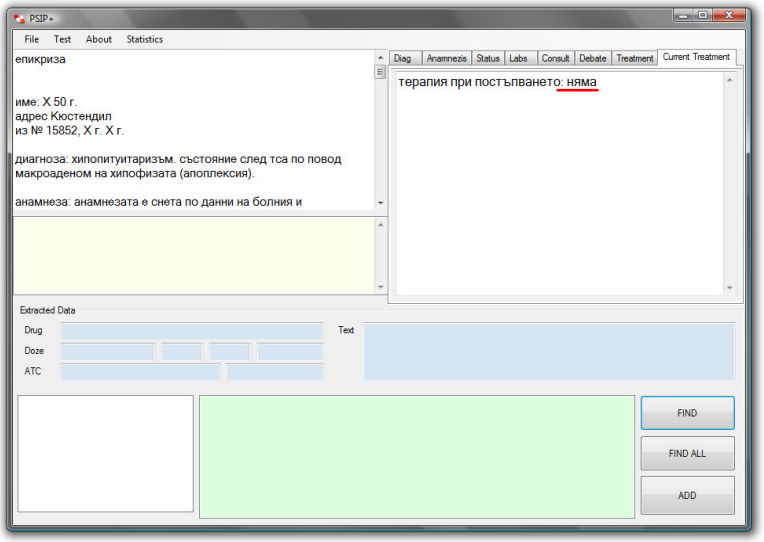


Fig. 6 Drug events - negation markers

In some cases even there is available "Current treatment" section in PR it contains negation marker.

In this example "няма" (No) causes association of empty drug events list for current treatment at hospital admission.

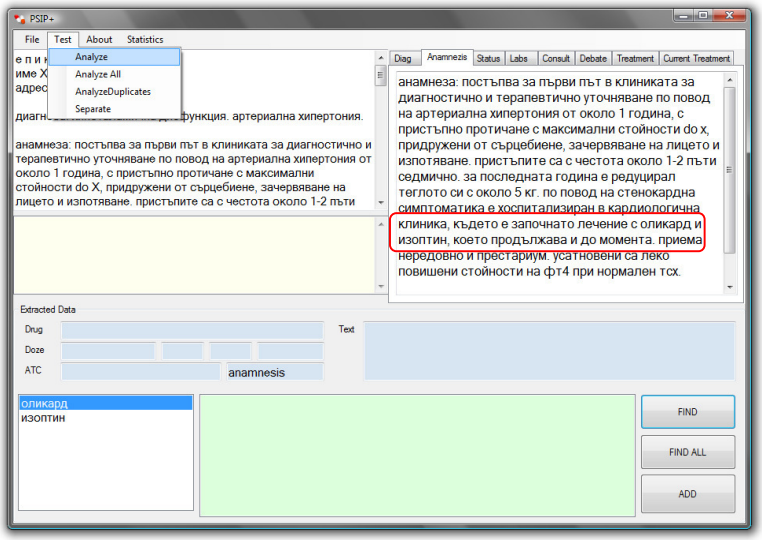


Fig. 7 Drug events - Anamnesis

In this case "Current treatment" section is not available and the system processes temporal events in "Anamnesis" section. In this example as current treatment is recognized the sentence "по повод на стенокардна симптоматика е хоспитализиран в кардиологична клиника, където е започнато лечение с оликард и изоптин, което продължава и до момента." *(….due to anginal symptoms was hospitalized in the cardiology clinic, where started treatment with with Olicard (Olicard 40 retard) and Isoptin, and which continues to date …)* Because for Olicard and Isoptin is not presented information for dosage we use N/A marker and further use DDD (default daily dosage).

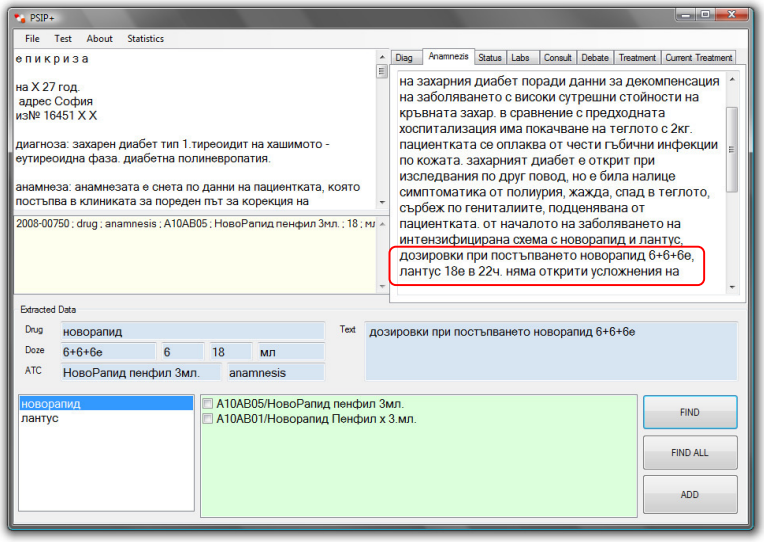


Fig. 8 Drug events - Anamnesis

In this case "Current treatment" section is not available and the system processes temporal events in "Anamnesis" section. In this example as current treatment is recognized information in the sentence "от началото на заболяването на интензифицирана схема с новорапид и лантус, дозировки при постъпването новорапид 6+6+6E, лантус 18E в 22ч." (Since the beginning of the decease with intensified regimen of treatment with Lantus and NovoRapid, dosages at hospital admission NovoRapid 6+6+6E , Lantus 18e at 10 pm.)

For the scheme 6+6+6E for NovoRapid is calculated daily dosage 18E.

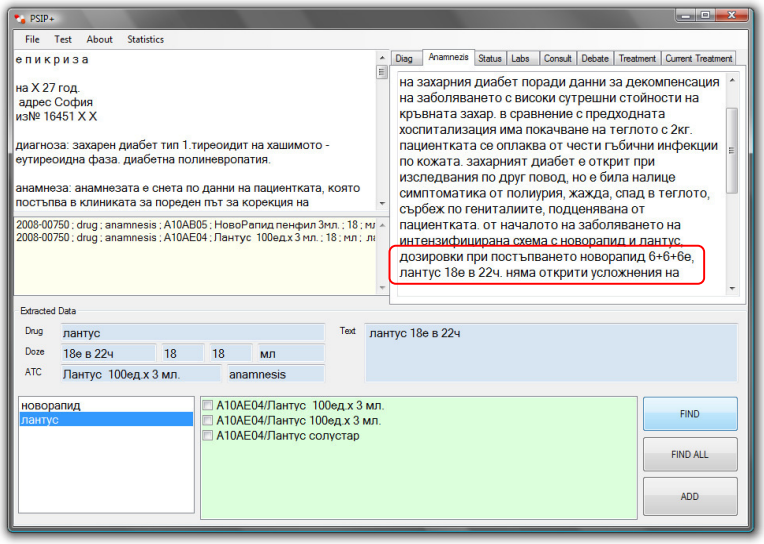
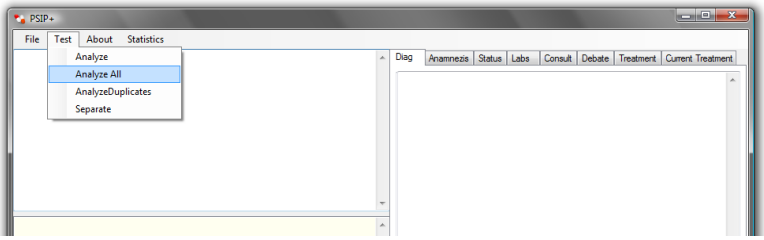


Fig. 9 Drug events - Anamnesis

From example in Fig.8 for Lantus is recognized daily dosage 18E. If some drug name is mentioned more than one time in the text, than using rules for temporal events analyses we choose the dosage for the closest to present time marker.




```
2008-00152 ; G02CB03 ; Достинекс ; 2 ; т ; лечение при постъпването: достинекс 2х1т/седм
2008-00152 ; H03AA01 ; Л тироксин 0.5 мг. ; 50 ; мкг ;  л- тироксин 50мкгр/дн
2008-00152 ; C08DB01 ; Дилтиазем ; 180 ; мг ;  дилтиазем 2х90мг
2008-00153 ; C08DB01 ; Дилтиазем ; 180 ; мг ; лечение при постъпването: дилтиазем 2х90 мг
2008-00153 ; C02AC01 ; Хлофазолин ; 1 ; таб ;  хлофазолин 2х1/2таб /поради предстоящите изследвания/
2008-00153 ; C01DA08 ; Изодинит 10мг х 60 ; 2 ; т ;  изодинит 2х1т
2008-00153 ; N05BA05 ; Транксен ; 1 ; т ;  транксен 1 т
2008-00155 ; B01AC04 ; Плавикс ; 1 ; т ;  плавикс- 1т
2008-00155 ; C07AB12 ; Небилет 5.45 мг.х 14 ; 2 ; т ;  небилет- 5мг- 2х1т
2008-00155 ; C09BA07 ; Цибадрекс ; 20 ; мг ; артериална хипертония- цибадрекс- 20мг
2008-00155 ; C10AB05 ; Липантил 200 мг.х 30 бр. ; 200 ; мг ;  липантил- 200мг
2008-00156 ; A10BA02 ; Метформин НИХВИ табл.х 30 бр. ; 1,5 ; т ;  метформин- 3х1/2т
```

Fig. 10 Drug events - Automatic mode

In automatic mode can be processed all PRs stored in the selected folder. The result file can be generated in CSV format, XML format or Excel format. Extracted information contains PR ID, ATC, drug name and drug pack, daily dosage, mode and scope of the recognized drug event from the narrative text.

3. Diagnoses Information Extraction

The Bulgarian hospitals are reimbursed by the National Insurance Fund via the "clinical pathways" scheme. When a patient is hospitalized, they often select from the Hospital Information System (HIS) menu one diagnosis which is sufficient for the association of the desired clinical pathway to the respective patient. Thus most of complementary diseases diagnosed by the USHATE medical experts are recorded in the personal history as free text. They are entered in the Discharge letter section *Diagnoses* as free text and some of them are only mentioned in the section *Discussion* of the discharge letter.

This developed an approach for automatic mapping of International Classification of Diseases 10th revision (ICD-10) [6] to diagnoses extracted from discharge letters. The proposed algorithms are designed for processing free text document in Bulgarian language.

Diseases are often described in the medical patient records as free text using terminology, phrases and paraphrases which differ significantly from the ICD disease description. than those used in ICD-10 classification. In this way the task of diseases recognition (which practically means e.g. assigning standardized ICD codes to diseases' names) is an important natural language processing (NLP) challenge [8].

To solve this task we are using the following resources provided in Bulgarian language [6]:

- ICD-10 in Excel format;
- Index of diseases and pathological states and their modifications (Fig. 11).

Падане
– на нокти L60.8
– преждевременно, на млечни (временни) зъби K00.6
Падане, на радиоактивни частици (неблагоприятно въздействие), НКД T66
Паднали сводове M21.4
Палатоплегия K13.7
Палатосхизис (*виж също* Цепка, небце) Q35.9
Палец – *виж също състояние*
– смучене на (проблем при дете) F98.8
Палец, hallux
– валгус (придобит) M20.1
– – вроден Q66.8

Панецефалит, подостър, склерозиращ A81.1
Паникулит (на) M79.3
– гръбначен стълб M54.0
– нодуларен, несупуративен M79.3
– при lupus erythematodes L93.2
– рецидивиращ M35.6
– сакрален M54.0
– шия M54.0
Панически(-а) (пристъп) (състояние) F41.0
– реакция на стрес (преходна) F43.0
Панкардит (остър) (хроничен) I51.8
– ревматичен I09.8
– – активен или остър I01.8
Панкреас, панкреатичен – *виж*

Fig. 11 Index of diagnoses and

The other obstacle is mixture of Latin and Bulgarian terminology used in free text diagnoses presentation. For some of Latin terms is used transliteration in Cyrillic.

This component works in three steps [1,6]: (i) shallow text analysis by regular expressions and patterns matching, (ii) searching disease names in the terminology resource bank - medical terminology dictionary, list of abbreviations rules and Latin – Cyrillic transliteration rules, and (iii) application of terminology binding rules manually added by experts. The regular expressions, applied at step (i) for shallow syntactic analysis, encode grammatical patterns of text phrases which describe medication events in the particular training corpus (of endocrinology patients treated at USHATE). These expressions are extracted semi-automatically from the training texts by machine learning techniques.

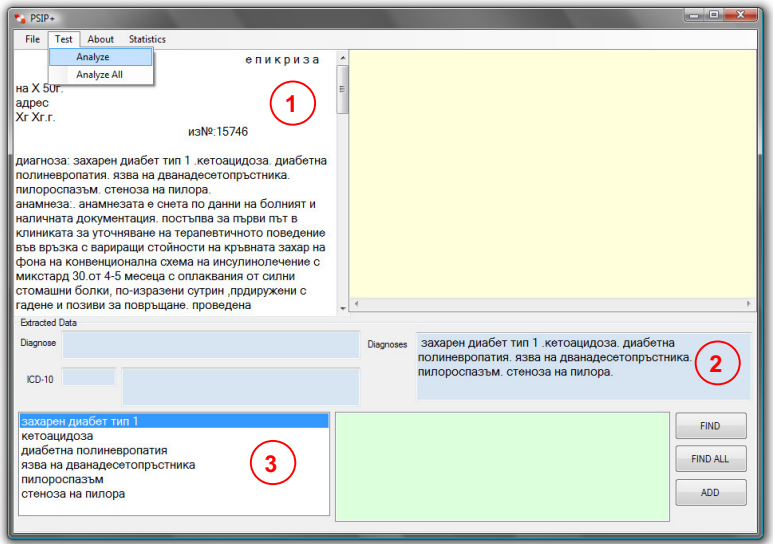


Fig. 12 Diagnoses manual analysis

Allows processing of a single PR stored as text file. The text of PR is opened in section (1). After opening the text file, PR is automatically separated on sections and the text from diagnoses section is displayed in section (2). After choosing "Analyze" function from menu bar the extracted text in section (2) is processed and automatically is generated list (2) with recognized diagnoses within the text.

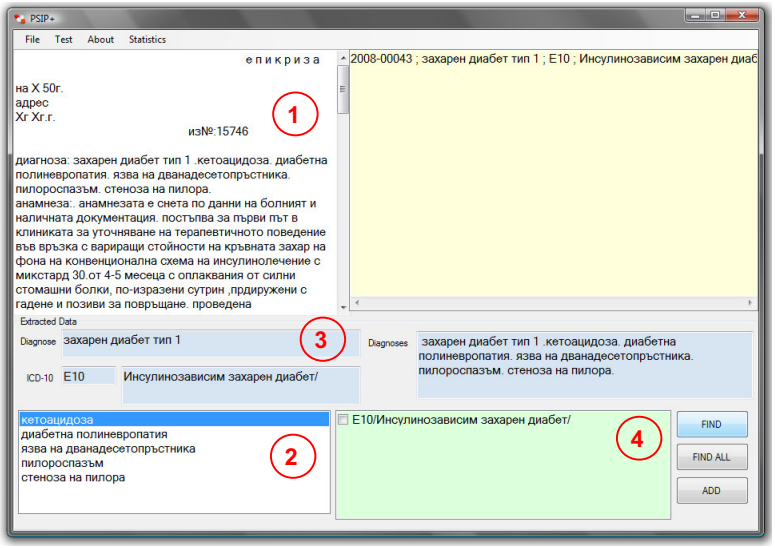


Fig. 13 ICD-10 codes assignment

Diagnoses included in the generated list (2) can be processed separately by using "Find" button and selecting diagnose from the list or as bunch using "Find All" button. After selection of diagnose from list (2) to be processed its name is automatically excluded from list (2) and displayed in section (3). In the current example the selected diagnose "захарен диабет тип 1" (Diabetes mellitus type 1) is displayed in sections (3). The system identifies possible ICD-10 codes assignments and displays them in list (4) – E10 Инсулинозависим захарен диабет (Insulin-dependent diabetes mellitus)

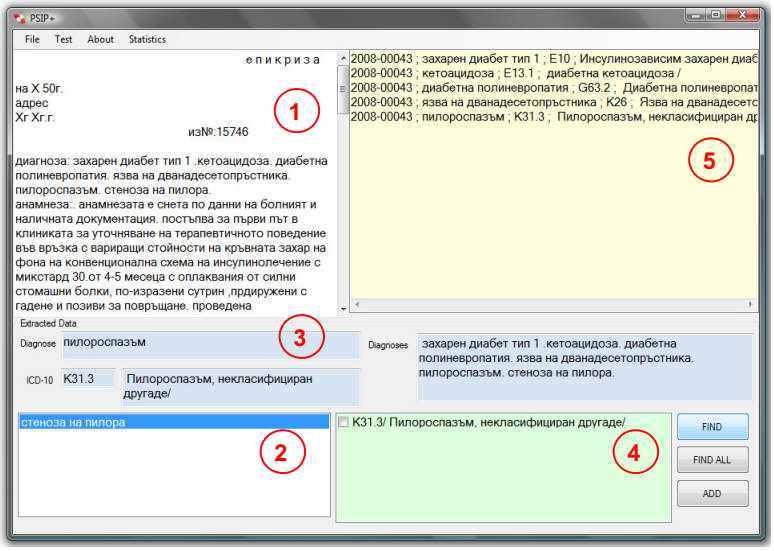


Fig. 14 ICD-10 codes assignment

The data for processed diagnoses from list (2) are displayed in list (5) for further storage in CVS format text file.

It is possible the system to identify more than one possible codes for assignment, in this case different options are displayed in list (4) in decreasing order of ranking. The most appropriate association is ranked first. There are two options – automatic assignment of ICD-10 code and manual assignment. If there are more than one options listed in (4), user can choose by checkbox the more appropriate code. In automatic mode the first ICD-10 code with highest rank is associated.

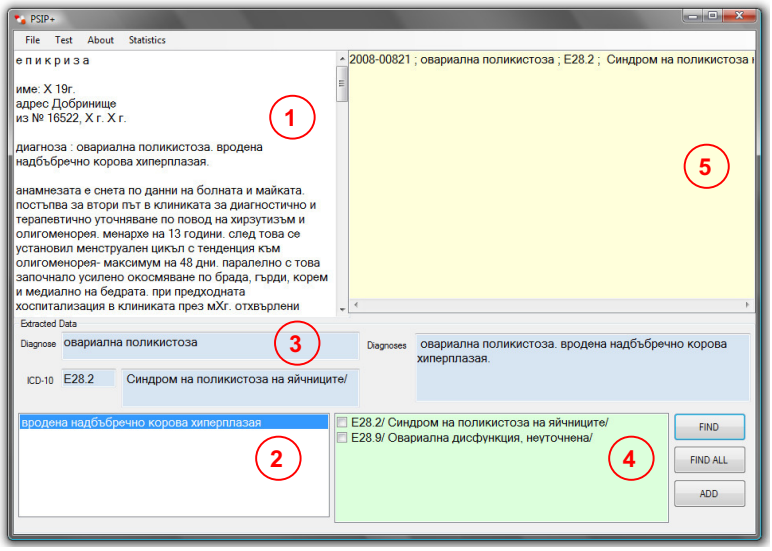


Fig. 15 Ranking

In this example diagnose is "овариална поликистоза" (Polycystic ovarian syndrome). For Latin term "овариална" (ovarian) (яйчници – in Bulgarian) in ICD-10 are recognized N83, Q50, C56, D27, E28 from 3 signs codes. We exclude C56 and D27 branches, because none of "neoplasm", "carcinoma in situ" and "melanoma" were presented. For further specification we search in 4 signs codes. For the second term "поликистоза" (Polycystic) we set highest rank to E28.2 which contains it, for the other codes we recognize Q50.3, N83.2 and E28.9 that refer to unspecified disorders.

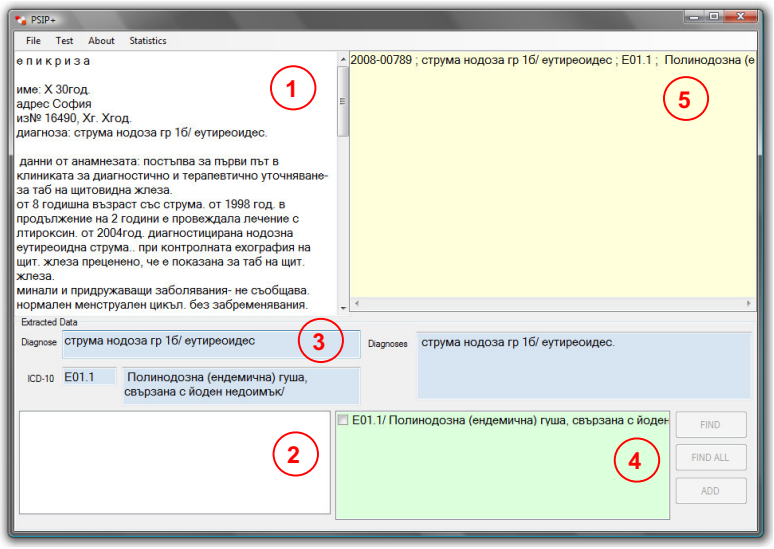


Fig. 16 Latin terminology processing

In this example the diagnose “струма нодоза гр 1б/еутироидес”(Struma Nodosa Euthyroides in Latin) (Euthyroid nodular goiter in English) is presented using latin terminology with transliteration. The first term “струма” corresponds to “гуша” (goiter) in Bulgarian language. In ICD-10 3 sign codes it corresponds to the cluster E00-E07.

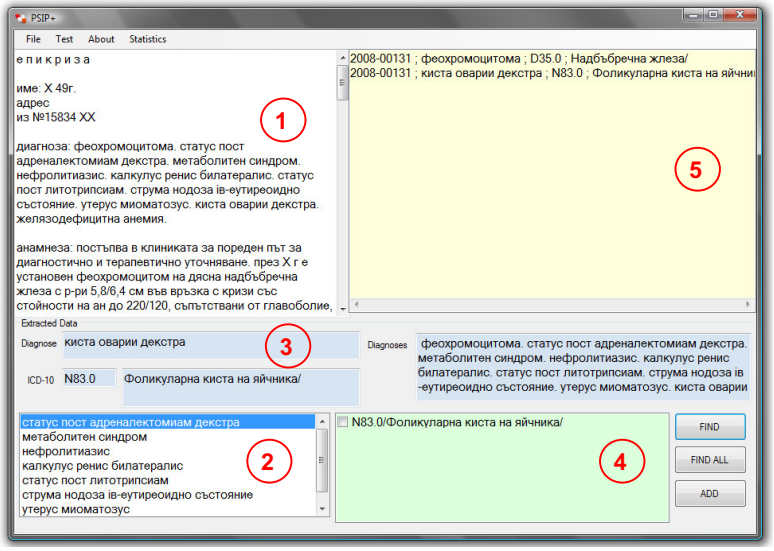


Fig. 17 Latin terminology processing

In this example the diagnose “феохромицитома” (pheochromocytoma) is presented using latin terminology with transliteration. This term corresponds to “Доброкачествено новообразувание на надбъбречна жлеза” (neoplasm of Adrenal gland) in Bulgarian language. In ICD-10 4 sign codes it corresponds to D35.0. The next diagnose “киста оварии декстра” (киста на яйчника – in Bulgarian, cyst of ovary – in English) () is processed similarly to example on Fig. 15 with assigned code N83.0 Фоликуларна киста на яйчника (N83.0 Follicular cyst of ovary). “декстра” in Latin means (дясна – in Bulgarian, Right – in English) is not considered in classification in this case

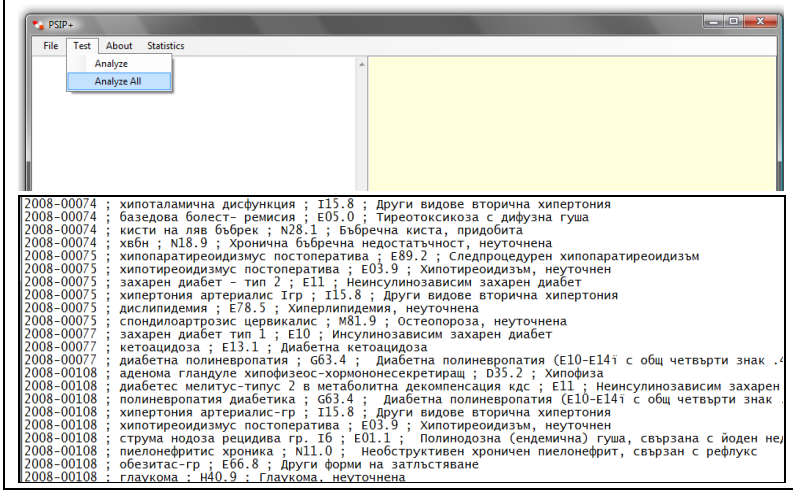


Fig. 18 Diagnoses automatic analyses

In automatic mode can be processed all PRs stored in the selected folder. The result file can be generated in CSV format, XML format or Excel format. Extracted information contains PR ID, diagnose name from the narrative text in PR, ICD-10 code, and diagnose name according to ICD-10 classification.

4. Evaluation Results and Discussion

In NLP the performance accuracy of text extraction procedures usually is measured by the *precision* (percentage of correctly extracted entities as a subset of all extracted entities), *recall* (percentage correctly extracted entities as a subset of all entities available in the corpus) and their harmonic mean *f-measure*: F=2*Precision*Recall/(Precision+Recall).

The experiments were made with a training corpus containing 1,300 PRs and the evaluation results are obtained using a test corpus, containing 6,200 PRs. In the test corpus there are 5,859 PRs with prescribed drugs during the hospitalization. The remaining 341 PRs concern patients hospitalized for clinical examinations only; these 341 PRs are excluded from the evaluation.

Evaluation results (Table 1 and Table 2) shows high percentage of success in drug name, drug dosage and diagnoses recognition in PRs texts.

Table 1 Extraction sensitivity according to the IE performance measures

	Precision	**Recall**	**F-Score**
Drug Name	97.28%	99.59%	98.42%
Dose	92.25%	95.51%	93.85%
Diagnoses	97.3 %	74.68*	84.5%

Table 2 Extraction sensitivity according to the IE performance measures

	Extracted entities from the PRs text
Drug Name	**160 892**
Diagnoses	**26 826**

The major reasons for incorrect recognition of medications are: misspelling errors, unrecognized drug events for allergies, incorrect detection of negation scope, drug events occurrence in other context and etc.

The incorrect assignment of ICD-10 codes for diagnoses is mainly due to: misspelling errors, unrecognized abbreviations, incorrect transliteration of Latin terminology and description of specific pathological states which is hard to classify according to ICD-10 even for humans.

Comparing with systems performing similar task in the PSIP project like French Multi-Terminology Indexer (F-MTI) [9], which indexes documentation in several health terminologies, we note that despite all complications in processing USHATE discharge letters the performance of our system is satisfactory. F-MTI is applied for automatic detection of Adverse Drug Events in discharge letters. The extraction of ATC codes from the free text of French discharge letters is performed with f-measure 88% when compared to the manual extraction; however, compared to the CPOE content, the f-measure is 49%. The main reason for better performance of our systems is that the discharge letters in French seem to have no predefined structure, which is available in Bulgaria that significantly helps to recognize events.

During the Third i2b2 Shared Task and Workshop "Challenges in Natural Language Processing for Clinical Data: Medication Extraction Challenge" [10] several semi- and un-supervised systems for medical information extraction were presented. They report result for machine learning base algorithms [11] high f-measure (91.40% for medication and 94.91% for dosage). Statistical hybrid methods that combine machine learning and rule-based modules [12] report f-measure for medication 89.9% and for dosage 93.6%. The other approaches for medication information recognition are mainly based on techniques like information extraction [13], rule-based [14], event driven [15] and semantic mining [16] and report similar performance.

For the second task – assignment of ICD-10 codes to diagnoses the results are comparable with recent systems as MIDAS (Medical Diagnosis Assistant) [17], SynDiKATe [18] with about 76% F-measure based on combination between text parsing and semantic information derivation from a Bayesian network, the system reported in [19] uses a hybrid approach combining example-based classification and a simple but robust classification algorithm (naive Bayes) with high performance over 22 millions PRs: f-measure 98,2%; for about 48% of the medical records at Mayo clinic, another 34% of the records are classified with f-measure 93,1%, and the remaining 18% of the records are classified with f-measure of 58,5%. The article [20] compares three machine learning methods on radiological reports and points out that the best f-measure is 77%.

5. Conclusion and Further Work

The paper presents software modules for PSIP+ project which supports the automatic extraction of medication information and diagnoses from PR texts

The Semantic mining modules are strictly oriented to Bulgarian language. The plans for their further development and application are connected primarily to Bulgarian local context. Future enhancements are planned for extension of the name and dosage recognition rules, to cope with certain specific exceptions and section filtering rules. The preliminary correction of spell errors and other kinds of typos will also increase the IE accuracy.

For diagnoses recognition task we plan improvement of rules for more precise code assignments.

6. Acknowledgements

The research tasks leading to these results have received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 216130 PSIP (Patient Safety through Intelligent Procedures in Medication).

7. References

[1] 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.

[2] 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/.

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

[4] *ATC drugs classification,* http://atc.thedrugsinfo.com/, last visited on 01/06/2011

[5] Bulgarian Drug Agency, *Drug List* at http://www.bda.bg/images/stories/documents/register/Mp.htm

[6] National Center of Health Information, http://www.nchi.government.bg/download.html

[7] Boytcheva, S. Assignment of ICD-10 Codes to Diagnoses in Hospital Patient Records in Bulgarian. In: Alfred, R., G. Angelova and H. Pfeiffer (Eds.). *Proc. of the Int. Workshop "Extraction of Structured Information from Texts in the Biomedical Domain" (ESIT-BioMed 2010*), ICCS-2010, Malaysia, Published by MIMOS BERHAD, ISBN 978-983-41371-3-7, July 2010, pp. 56-66.

[8] Demner-Fushman, D., W. Chapman and C. McDonald. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics, Volume 42, Issue 5, October 2009,* (2009), 760-772.

[9] 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? In *Studies in Health Technology and Informatics*, *Proceedings of the 13th World Congress on Medical Informatics*, Cape Town, South Africa, Volume 160, Number pt 1, (2010), 1025-1029.

[10] Third i2b2 Shared-Task and Workshop "Challenges in Natural Language Processing for Clinical Data: Medication Extraction Challenge", https://www.i2b2.org/NLP/Medication/.

[11] Patrick, J. and M. Li. A Cascade Approach to Extracting Medication Events. *In: Proc. Australian Language Technology Workshop (ALTA),* (2009), 99-103.

[12] 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.

[13] Meystre, S. M., G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research, *IMIA Yearbook 2008: Access to Health Information*, Vol. 1 (2008), 128-144.

[14] Evans, D. A., N. D. Brownlowt, W. R. Hersh, and E. M. Campbell. Automating Concept Identification in the Electronic Medical Record: An Experiment in Extracting Dosage Information. *AMIA 1996 Symposium Proceedings*, (1996), 388-392.

[15] Miura, Y., E. Aramaki, T. Ohkuma, M. Tonoike, D. Sugihara, H. Masuichi, and K. Ohe, Adverse-Effect Relations Extraction from Massive Clinical Records, *Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, Coling 2010 Beijing, 2010, 75-83.

[16] Chazard E., C. Preda et al., *Deliverable D2.3 - Results of data & semantic mining, PSIP Project*, URL: https://www.psip-project.eu/2010

[17] Sotelsek-Margalef, A. and J. Villena-Román. MIDAS: An Information-Extraction Approach to Medical Text Classification (MIDAS: Un enfoque de extracción de información para la clasificación de texto médico), Procesamiento del lenguaje Natural n. 41, 2008, pp. 97-104.

[18] Hahn, U., K. Schnattinger and K. Markó. *Wissensbasiertes Text-Mining mit SynDiKATe (Knowledge based text mining with SynDiKATe)*. Künstliche Intelligenz, vol. 2, 2002.

[19] Pakhomov, S., J. Buntrock and C. G. Chute. Automating the assignment of diagnosis codes *to patient encounters*, Journal of American Medical Informatics Association, 13, 2006, pp. 516-52.

[20] Coffman, A. and N. Wharton. Clinical Natural Language Processing: Auto-Assigning ICD- *9 Codes*. Overview of the Computational Medicine Center's 2007 Medical Natural Language Processing Challenge. Available online at http://courses.ischool.berkeley.edu/i256/f09/Final%20Projects%20write-ups/coffman_wharton_project_final.pdf