# Shallow Medication Extraction from Hospital Patient Records

Svetla Boytcheva,
*Institute of Information and Communication Technologies,*
*Bulgarian Academy of Sciences,*
*25A Acad. G. Bonchev Str. Sofia, Bulgaria*
*and University of Library Studies and Information Technologies*
*svetla.boytcheva@gmail.com*

**Abstract.** This paper presents methods for shallow Information Extraction from the free text zones of hospital Patient Records (PRs) in Bulgarian language. We extract automatically information about drug names, dosage, modes and frequency. The prototype of the system was used for experiments with 6200 records and performs extraction with more than 90% accuracy.

**Keywords.** Information extraction, automatic patient record processing, patient treatment information

## Introduction

Huge amount of clinical narratives are produced allover the world every day; free text is convenient for expressing details about patients but is difficult for automatic processing. One of the most important challenges in biomedical informatics nowadays is to find efficient methods for information extraction from unstructured texts. The main difficulties are due to the specific medical language: large amount of terms, variety of expressions describing clinical events, rich temporal information, negations of various kinds, much explicit and tacit knowledge needed for proper interpretation and so on. In particular the Bulgarian medical texts contain a specific mixture of terminology in Latin, Cyrillic and Latin terms transcribed with Cyrillic letters. The lack of nomenclatures, corpora, and electronic dictionaries for medical terminology in Bulgarian language makes the task of automatic text processing even harder.

We have developed automatic procedures 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 to each medication event. We deal with hospital PRs which are anonymised by the hospital information system of the University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev" (USHATE) at Medical University – Sofia.

This paper is organised as follows. Section 1 presents an overview of related work. Section 2 describes the resource bank used in our prototype. Section 3 discusses the system architecture and some examples; it presents our approach and the main problems that need to be solved. Section 4 summarises the experiment results and the evaluation. Section 5 contains the conclusion and sketches ideas for further work.

## 1. Background and Related Work

Natural language processing (NLP) is viewed as the most promising technology for capturing information from free text documents. Here we briefly overview the major NLP approaches with focus of automatic identification of drugs and adverse drug events in the text. During the Third i2b2 Shared Task and Workshop "Challenges in Natural Language Processing for Clinical Data: Medication Extraction Challenge" [1] several semi- and un-supervised systems for medical information extraction were presented, e.g. [2]. The most popular approaches for solving this task are:

- ***Information Extraction*** (IE) - simple pattern matching techniques and partial shallow analysis are widely used in biomedical text processing, see a recent review of systems which extract information from textual documents in the electronic health records [3].
- ***Rule-based methods*** recognise well the regular configurations of text entities. For instance, the NLP system CLARIT extracts drug-dosage information from clinical narratives using pattern matching based on regular expressions [4]. Text analysis is accomplished in five steps: tokenization, stemming, syntactic category assignment, semantic category assignment and pattern matching.
- ***Machine learning*** is another popular NLP technique. For instance, the article [5] presents a cascade approach for extracting medication information. The implemented system recognises medication events by combining machine learning and a rule-based approach. Two machine learners were used, namely the Conditional Random Field (CRF) and Support Vector Machine (SVM). The authors report high recall (91.44% for medication and 93.49% for dosage), high precision (91.35% for medication and for 96.36% dosage) and correspondingly high f-measure (91.40% for medication and 94.91% for dosage). Another SVM-based named entity recognition system for extraction of medication related entities achieves best f-score of 90.05% [6].
- ***Statistical hybrid methods*** combine machine learning and rule-based modules. The article [7] presents a hybrid system performing medication information extraction. With only a handful of template-filling rules, the system's core is a cascade of statistical classifiers for field detection. This system did not participate in the i2b2 Challenge but it achieves good results that match the top i2b2 systems: recall for medication 88.5% and for dosage 90.8%; precision for medication 91.2% and for dosage 96.6%; f-measure for medication 89.9% and for dosage 93.6%.
- ***Event driven approaches***: the extraction of adverse drug events and effect relations from clinical records is presented in [8]. The authors propose a method to extract adverse–effect relations using a machine learning technique with dependency features.
- ***Semantic mining*** comprises a set of ontology-based techniques which extract relevant information from medical letters and reports, using the main health terminologies [9]. Semantic mining applies NLP to capture information which is not included or is missing in the Hospital Information Systems and CPOE databases. Semantic mining provides for each medical letter or report relevant terms from different terminologies with their meaning and relations between them. Semantic Mining is closely related to various Natural Language Processing tools, therefore it addresses documents in specific languages.

The evaluation results cited above show that no contemporary NLP system provides extraction with 100% precision and recall. However, despite all difficulties to process automatically the narrative texts in the medical domain, the interest in the development of fundamental and applied NLP methods for medical text analysis is constantly growing. This is due to the fact that NLP is viewed as the only means for (partial) automatic understanding of medical documents [10]. Comparing the methods listed in this section we see that CRF delivers better results than the Rule-based approach, and the latter performs better than SVM.

## 2. Resource Bank

Unfortunately the presented IE techniques cannot be directly adapted to our project, because we deal with documents in Bulgarian and major language-processing activities start from scratch. First we need to cope with the morphological variants (drug names might occur in various wordforms due to the inflectional Bulgarian language). Phrasal patterns are acquired manually, to enable shallow sentence analysis by pattern matching with cascading applications of regular expressions. We partly use available linguistic resources but they support extraction of diagnoses and patient status [11]. Thus the medication IE started by the development of lexicons and training corpora.

The list of registered drugs in Bulgaria is provided by the Bulgarian Drug Agency [12]; it contains about 4000 drug names and their ATC codes. The main reference list uses the Latin drug names and the Bulgarian translations are provided in additional pdf-files. However, the patient records in USHATE use mostly Bulgarian drug names, so we needed to compile a Bulgarian lexicon of drug names. The Hospital Pharmacy (HP) supports names in two languages (see HP entries at Fig. 1): ATC code, drug names in Bulgarian and English, pharmacy code, dose, etc. Currently the HP operates with 1537 medications because USHATE is specialised mostly for treatment of diabetic patients.

| UNIQUE ID | ATC code | Pharmacy code | Drug Name in Bulgarian | Pharmacy Unit | Dose | Dose Unit | Drug Name in EN |
|---|---|---|---|---|---|---|---|
| 52 | A10AB01 | 01000054 | Инсулин актрапид МС | амп | 300,000 | I.U. | Insulin actrapid MC |
| 53 | A10AB01 | 01000055 | Инсулин актрапид НМ | амп | 300,000 | I.U. | Insulin Actrapid HM |
| 448 | A10AB01 | 01000560 | Хумулин R 40E 10Ммл | бр | 400,000 | I.U. | Humulin R 40 |
| 455 | A10AB01 | 01000567 | Актрапид 40E 10мл | бр | 400,000 | I.U. | Actrapid 40 |

**Figure 1.** Excerpts of drugs-related records in the USHATE Hospital Pharmacy

By matching lists of Bulgarian drug names, compiled from various sources including informal public sites in Internet, we have found 304 drugs that are mentioned in the USHATE hospital PRs but are not prescribed via the Hospital Pharmacy. These drugs occur in the free PR texts because they are taken by the patients to cure additional (chronic) illnesses while USHATE HPs contain records of drugs curing the diabetes. For instance, hypertony is a typical accompanying disease, and normally the patients arrive to USHATE bringing the medications prescribed by their GPs. In this way our present system processes 1841 drug names in Bulgarian and their ATC codes.

The Defined Daily Dose (DDD), associated to the ATC-classification, helps to assign default dosages when they are not explicitly mentioned in the PR texts. Lists of measurement units (both in English and Bulgarian) and various abbreviations support the recognition of text fragments discussing medication events. Our resources also contain several regular expressions and rules for (phrasal) pattern matching.

## 3. System Architecture

The length of PR texts in Bulgarian hospitals is usually 2-3 pages. The document is organised into the following sections: *(i)* personal details; *(ii)* diagnoses of the leading and accompanying diseases; *(iii)* anamnesis (personal medical history), including current complains, past diseases, family medical history, allergies, risk factors, and medical examiners comments; *(iv)* patient status, including results from physical examination; *(v)* laboratory and other tests findings; *(vi)* medical examiners comments; *(vii)* discussion; *(viii)* treatment; *(ix)* recommendations. Medication information is contained in sections *(iii)* anamnesis, *(vii)* discussion, *(viii)* treatment, and *(ix)* recommendations. Practically we need to process almost all text fragments in the PR.

Figure 2 presents the typical occurrences of medication descriptions in the PR texts. There are more than 50 different patterns for matching text units discussing medication name, dosage and frequency; four patterns are illustrated at Fig. 2.

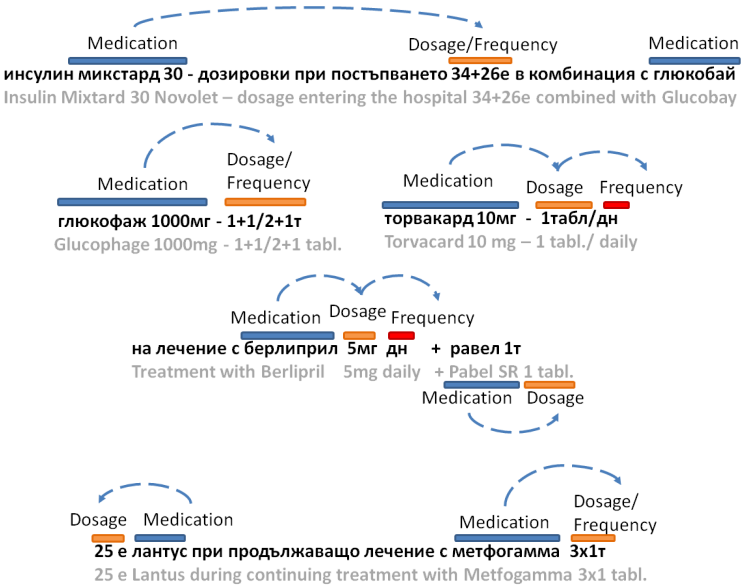


**Figure 2.** Sample patterns for recognition of text units expressing medication, dosage and frequency

The architecture of the system, which extracts medication information from hospital PRs in Bulgarian language, is shown in Figure 3. It contains eight modules:

*Sections splitting* - this module separates the PR texts into standardised sections. The splitting is not trivial due to varying section names, various abbreviations used to name the sections, missing sections in the PRs or missing section names, and swapped sections. The PRs summarise major patients' diseases and their treatment; the system searches medications in the whole PRs and the correct section splitting enables capturing of some temporal relations: the current treatment is presented in the anamnesis – esp. medical examiners comments, discussion, treatment and recommendations. The information about accompanying diseases and drugs which are not prescribed by the Hospital Pharmacy is given in the anamnesis section, as well as

the discussion of allergies and risk factors, so it is important to fix the section boundaries correctly.

*Sentence splitting* - this module separates the sentences in each PR section which facilitates the further text analysis. Missing delimiters are the major difficulties in this task. Usually the PR sentences end with a period, a colon, or the end of the line, but due to several abbreviations and formatting styles additional rules for sentence splitting are needed.

*Tokenization* - the input PR text is split into words, digital literals and punctuation.

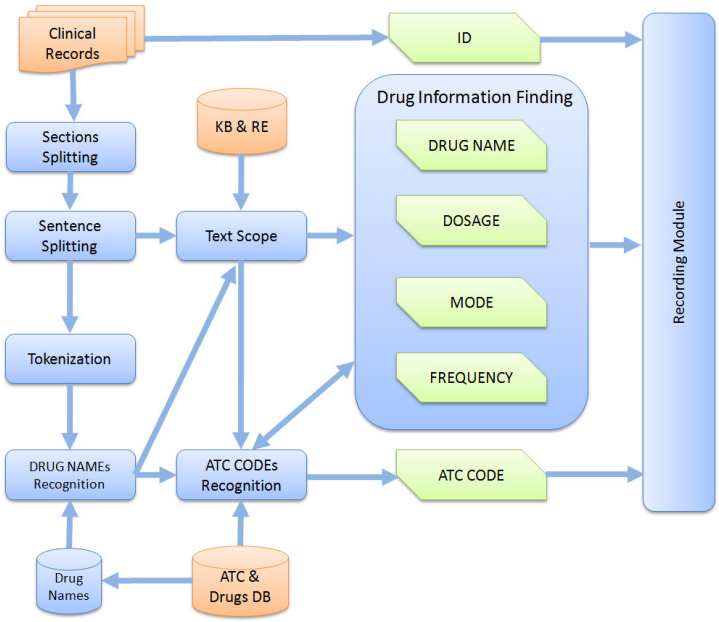


**Figure 3. System Architecture: main components, resources, and workflow** Drug Names?

*Drug names recognition* – this module matches the 1841 items of the drug list to the words in the PR sections. Some drug names occur several times in the text. The resulting list contains PR drug names without duplications. The main difficulties in this task are due to the fact that *(i)* many drugs in the list have names longer than one word and *(ii)* there is a huge variety of drug descriptions in PR text: names given in Latin or in Latin transcribed with Cyrillic letters; names given by abbreviations; short names or generic descriptions given instead of full brand name. For instance: "*витамин с*" (*vitamin C*) in the PR text has to be recognised as the brand name "*Вит Ц 100мг 40бр*"; "*хумулин н*" (*Humulin N*) in PR text has to be matched to "*Хумулин N*"; "*лтироксин*" (*L-Thyroxin*) or "*л тироксин*" in the PR text is actually "*Л-тироксин 50 мг.*"; "*апидра солостар*" (*Apidra*) or "*апидра*" in the PR text has to be recognised as the brand name "*Апидра Солустар*". The algorithm first tries to match the full names, if this fails different matches of name variations are tried and finally, skipping or swapping some of the words is tried.

*Text scoping* – this module finds the text fragment which contains the actual information about dosage and frequency for each drug name. We assume that the last

drug name's occurrence contains the actual treatment information. Sometimes the dosage and frequency are mention together with the previous drug name occurrences, and the last one contains only information that the previously prescribed dosage needs to be increased, decreased, doubled or remain unchanged. In this case the system finds the previous occurrence of the same drug name and captures the dosage from there, and then refines the dosage and frequency information according to last occurrence. However, as it was shown in Fig. 2, the scope of the text conveying drug names and dosage can be quite wide; this text can also contain elliptical constructions with other drugs with equivalent dosage and frequency. The text scope is determined by a cascade approach for regular expressions matching onto the PR text. The text scoping algorithm uses names of measures and a lexicon of abbreviations for dosage units' detection.

*Drug information finding* – this module captures information about drug name, dosage, mode and frequency from the scoped text using regular expressions. If it succeeds, the result is given to the next module. If the dosage, frequency or mode/route are not recognised (because explicit details are missing in more than 30% of the PR descriptions), the drug name is passed to the next module for assignment of an ATC code and then the DDD is selected as a default value.

*ATC code recognition* – after the identification of drugs in the text the system finds the appropriate brand name and the corresponding ATC code. For instance, in the case of "*еналаприл 2х20 мг*" (*Enalapril*) there are two options: "*Еналаприл 10 мг.х 30 бр.*" (*Enalapril tabl. 10 mg x 30)* and "*Еналаприл т. 20 мг.х 30 бр*." (*Enalapril tabl. 20 mg x 30)* with the same ATC code C09AA02. According to the dosage 20 mg the system chooses "*Еналаприл т.20 мг.х 30 бр.*" (*Enalapril tabl. 20 mg x 30)*. If no information about the dosage is available the algorithm chooses the first ATC code from the list.

*Recording module* – this module collects all data extracted by the previous modules and saves them in different formats – XML, ASCII or MS Excel table.

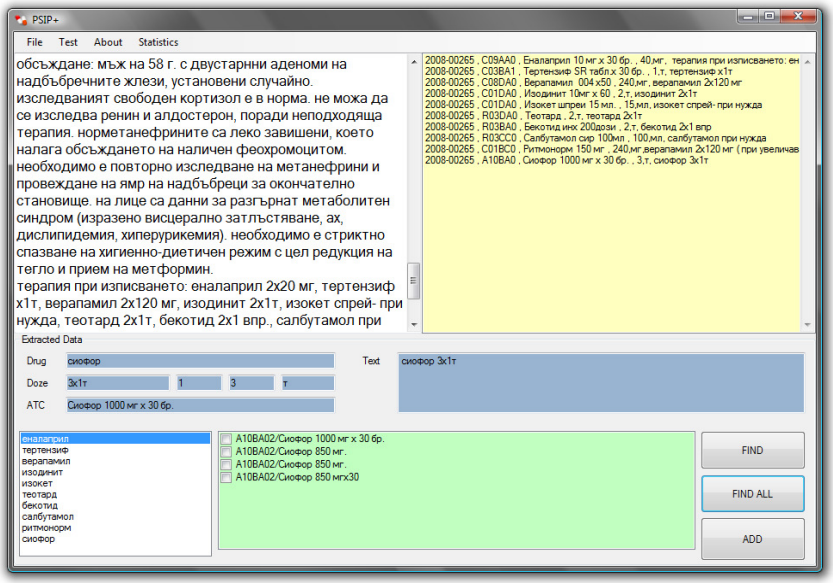


**Figure 4.** The user interface presenting the extracted medication data from a particular Patient Record

The system presented here can process PRs in *(i)* automatic mode – analysing all PRs from a chosen folder and producing a file with the extracted medication data and *(ii)* single mode – analysing PRs separately and presenting the results at the user interface. A sample from a single-mode analysis of one PR is shown at the screenshot in Figure 4. The PR contains information about 10 drugs; the system is ready to propose their ATC codes, dosage, mode and frequency. The last processed drug name is "*сиофор*" (*Siofor*); an ATC code and brand name from 4 options is chosen. In this way the user can test the system and evaluate its performance.

## 4. Evaluation Results and Discussion

The experiments were made with a training corpus containing 1300 PRs and the evaluation results are obtained using a test corpus, containing 6200 PRs. In the test corpus there are 5859 PRs with prescribed drugs during the hospitalisation. The remaining 341 PRs concern patients hospitalised for clinical examinations only; these 341 PRs are excluded from the evaluation.

Figure 5 shows the number of drugs taken by patients during their hospitalisation in USHATE. The maximal number of drugs is 27, the minimal number is 1 and the average number of drugs per patient is 5.43. Most often the patients take 2-4 drugs.

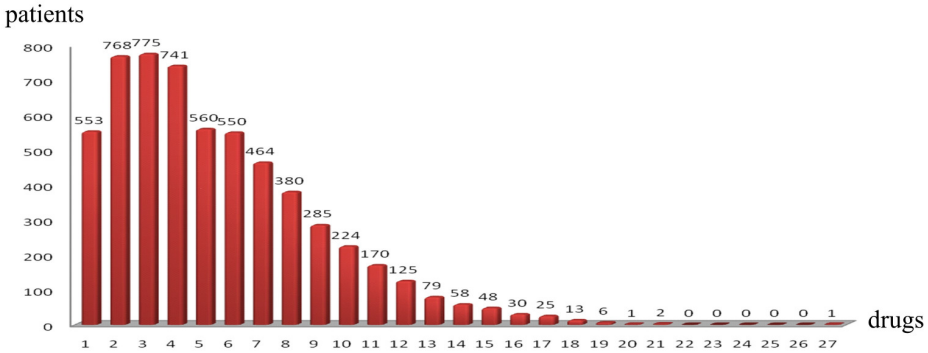


**Figure 5.** Number of drugs per patient.

The performance accuracy 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 $F=2*Precision*Recall/(Precision+Recall)$. The evaluation figures presented here summarise the IE performance for extraction of 667 different drugs (brand names) which were juxtaposed 346 different ATC codes. Evaluation results (Table 1 & Table 2) shows high percentage of success in drug name recognition in PRs texts. False negatives in Table 1 are mainly due to misspelling or too strict rules in the algorithm for recognition of drug names used in different context. False positives are mainly caused by some negation detection. We consider the negated descriptions as one expressing, following a study of negative forms in Bulgarian medical patient texts [13]. The true positive percentage is very high for drug names (30987 true positive out of 31853 records extracted from the test corpus, Table 1).

**Table 1.** Number of extracted medication events in 5859 PRs

| | All extracted | True positive | False positive | False negative |
|---|---|---|---|---|
| Drug Name | 31853 | 30987 | 836 | 127 |
| Dose | 26827 | 24750 | 2077 | 1163 |

**Table 2.** 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% | 93,8%5 |

Below we discuss the major reasons for incorrect recognition. Errors come from:

*(i)* Misspelling of drug names, such as "*лтироксин*" (*LThyroxin*) or "*л тироксин*" (*L Thyroxin*) instead "*Л-тироксин"* (*L-Thyroxin*)

*(ii)* Drug names occurring in the contexts of other descriptions, such as:

- Diagnosis – many drugs names participate in diseases' names, e.g. "*дефицит на витамин д*" (*Vitamin D deficiency*);
- Examination results – such as Calcium, Kalium, e.g. "*лечение с рокалтрол 2 капс, заедно с 1800 мг. Калций; на този фон серумният калций се е движил около долната граница на нормата*." (*Treatment with Rocaltrol 2 capsules, combined with 1800 mg Calcium; the serum Calcium level was closer to the lower range of the norm*);
- Hormones – such as Testosteron and Progesterone, e.g. "*hormone replacement therapy with high doses of estrogen (Estrofem 2mg), combined with progesterone preparation (Duphaston)"*.

*(iii)* Undetected descriptions of drug allergies –we have found 391 unrecognised cases, among them 316 for allergies, 25 for sensibility, 50 for intolerance and 36 for side effects;

*(iv)* Drug treatment described by (exclusive) *OR* – we have found about 30 cases of incorrect recognition in such kind of phrases, e.g. "*in case of deterioration the treatment should be replaced by Glucobay 3x100 or Amaryl 2mg*" and "*therapy: Tenaxum or Physiotens under the control of the blood pressure*"). In these cases both drugs are recognised and inserted in the resulting extracted records.

*(v)* Negations and temporally-interconnected events of various kinds:

- Undetected descriptions of canceled medication events – we have found 205 incorrect cases where the extracted drugs need to be excluded from the resulting records, e.g. "*the therapy with biguanides preparation (Glucophage) was stopped and Gliper was replaced by Diaprel"*;
- Undetected descriptions of changes or replacements in therapy – we have found 234 unrecognised phrases. In this case both drugs are extracted and the previous one is not deleted from the result records. For instance, this happens in the analysis of the sentence *"taking into consideration the pregnancy desire, to continue with Bromocriptine 2x1 tabl. daily without replacing the treatment by Dostinex*" as well as with *"2 weeks before entering the hospital replace Concor by Isoptin 2x120mg in order to examine the patient again*"*).*
- Undetected descriptions of insufficient treatment effect and change of therapy – we have found 6 cases of wrong extraction. For instance: "*due to the unsatisfactory effect of the treatment with Vigantol up to 20 drops per week the patient starts treatment with Rocaltrol 2x1 drops daily"*).

As seen at Table 1 and Table 2, the dosage recognition is less successful than the recognition of drug names. About 30% of the medication events in the test corpus were described without any dosage, e.g. "*to continue the treatment with Flarex and Azopt in the eyes*". Lack of explicit descriptions occurs mostly for treatment of accompanying diseases (because the attention of USHATE's medical expert is focused on the specialised hospital treatment, disregarding drugs that are prescribed by other clinicians beforehand). After applying the recognition algorithm and using the default DDD dosage, the number of records lacking dosage was reduced to 5026 or 15.7% in the final result containing 31853 records. For the PRs with explicitly declared dosages, the main sources of errors are the following ones:

- Mismatch between the PR text and the content of the respective Hospital Pharmacy/ATC values – for instance, "*C07AB0 / Atenolol 50 mg x 30*" in the Hospital Pharmacy and "*Atenolol 2x25 mg*" in the PR text. In this case the system recalculates the dosage according to the closest Pharmacy/ATC value;
- Unfixed dosage – for instance "*recommend treatment with Metformin from 3x850mg to 3x1000 mg / daily under control of the blood sugar profile*";
- Ambiguous dosage – "*лечението със сиофор 3х1т*" (*treatment with Siofor 3x1 tabl.*) but in Pharmacy we have "*Сиофор 1000 мг х 30 бр*" (*Siofor 1000 mg*) " *Сиофор 850 мг.*" (*Siofor 850 mg*)
- Partial or incomplete information about the therapy scheme or mixing dosage as part of the brand name, e.g. "*Сиофор 850 мг.*" (*Siofor 850 mg*) etc.

Despite all complications listed here, the precision and recall in the automatic recognition of drug dosage are relatively high as well (see Table 2). At present we complete the evaluation of the extraction procedures which recognise drug mode/route and frequency. Our present results are comparable to the performance of advanced systems such as MedEx [14]. Please note that we try to address negative statements, elliptical constructions, typical conjunctive phrases, simple inferences concerning temporal constraints and finally aim at the assignment of the drug ACT code to the extracted medication events, which additionally complicates the extraction algorithm.

## 5. Conclusion and Further Work

The system presented in this article was developed and applied in the PSIP project for the preparation of an experimental USHATE's repository for PSIP validation. 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. This system is a pilot prototype performing extraction of drugs and medication events from Bulgarian medical texts. The promising results support the claim that the Information Extraction approach is helpful for the obtaining of specific medication information from free patient record texts. The performance cannot be directly compared with other results reported in the literature, because of the language specific analysis techniques and the specific hospital personal records in Bulgarian language, but nevertheless the accuracy is relatively very high.

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.

## 6. Acknowledgements

This work is supported by the project PSIP (Patient Safety through Intelligent Procedures in medication), funded by the EC FP7 ICT grant 216130.

## References

[1] *Third i2b2 Shared-Task and Workshop "Challenges in Natural Language Processing for Clinical Data: Medication Extraction Challenge"*, https://www.i2b2.org/NLP/Medication/, last visited 03/01/2011

[2] Mork, J. G., O. Bodenreider, D. Demner-Fushman, R. I. Doğan, F.-M. Lang, Z. Lu, A. Névéol, L. Peters, S. E. Shooshan, and A. R. Aronson, Extracting Rx Information from Clinical Narrative, *Journal of the American Medical Informatics Association JAMIA* 17 (2010), 536-539.

[3] 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.

[4] 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.

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

[6] Doan S. and H. Xu, Recognizing Medication related Entities in Hospital Discharge Summaries using Support Vector Machine, *In Proc. of Coling 2010: Poster Volume, Beijing*, (2010), 259–266.

[7] 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.

[8] 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.

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

[10] 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.

[11] Boytcheva S., I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev and N. Dimitrova. Obtaining Status Descriptions via Automatic Analysis of Hospital Patient Records. In: V. Fomichov (Ed.), *Informatica, Int. J. of Computing and Informatics*, *Sp. Issue on Semantic IT,* vol. 34 (2010), 269–278.

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

[13] Boytcheva, S., A. Strupchanska, E. Paskaleva, and D. Tcharaktchiev, Some Aspects of Negation Processing in Electronic Health Records. *In Proc. of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries*, Borovets, Bulgaria, (2005), 1-8.

[14] 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. *Journal of the American Medical Informatics Association JAMIA* 17 (2010), 19-24.