Automatic Matching of ICD-10 codes to Diagnoses in Discharge Letters

Svetla Boytcheva
State University of Library Studies and Information Technologies,
119, Tzarigradsko Shosse, 1784 Sofia, Bulgaria
svetla.boytcheva@gmail.com

Abstract

This paper presents an approach for automatic mapping of International Classification of Diseases 10th revision (ICD-10) codes to diagnoses extracted from discharge letters. The proposed algorithm is designed for processing free text documents in Bulgarian language. Diseases are often described in the medical patient records as free text using terminology, phrases and paraphrases which differ significantly from 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. The approach is based on multiclass Support Vector Machines method, where each ICD-10 4 character classification code is considered as single class. The problem is reduced to multiple binary classifiers and classification is done by a max-wins voting strategy.

1 Introduction

The nomenclature ICD or ICD CM (International Classification of Diseases with Clinical Modification), supported by WHO (World Health Organization) [1], is translated to many languages and serves as the main source for diagnoses definition.

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 medical experts are recorded in the personal history as free text. To describe diseases as free text in the medical patient records (PRs) usually is used different terminology than those used in ICD-10 classification in order to express more specific and detailed information concerning particular disorder or using paraphrases, which usually are not available in general classification. For instance, in some diagnoses is specified the stage "затлъстяване I степен" (Stage 1 Obesity), the specific location "катаракта на ляво око" (left eye cataract) etc.

Thus the task of diagnoses recognition from free-text discharge letters and assignment of standardized ICD codes to diseases' names is an important natural language processing (NLP) challenge [2].

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 [3]. PRs contain the following sections: *(i)* personal data; *(ii)* diagnoses; *(iii)* anamnesis; *(iv)* patient status; *(v)* lab data; *(vi)* medical examiners comments; *(vii)* discussion; *(viii)* treatment; and *(ix)* recommendations. Most of the diagnoses are entered in the discharge letter section *Diagnoses* as free text and some of them are only mentioned in the *Discussion* or *Medical examiners comments.*

In this paper we present an approach based on multiclass Support Vector Machines (SVM) for automatic diagnoses recognition from free-text PRs and assignment the ICD-10 codes to them.

The paper is organized as follows: Section 2 overviews related work, Section 3 describes recourse bank, Section 4 presents the method, system architecture and some examples, Section 5

Proceedings of the Workshop on Biomedical Natural Language Processing, pages 11–18,
Hissar, Bulgaria, 15 September 2011.

discusses evaluation and results and Section 6 sketches further work and conclusion.

2 Related Work

The application of natural language processing methods to clinical free-text is of growing interest for both health care practitioners and academic researchers. Unfortunately there is no international standard for discharge letters presentation. Another main difficulty in such texts is medical terminology - German, English and French medical terminology mainly is based on domesticated terms, but still some Latin terms are used and some of them are modified by preserving Latin root and using domestic ending

There are no systems dealing with clinical texts in Bulgarian. Thus we will overview some of the recent results achieved mainly for processing discharge letters in English [6, 8, 9], German [7] and French.

Several methods dealing with this problem were presented on 2007 Computational Medicine Challenge where about 50 participants submitted results [4]. The main goal was to create and train computational intelligence algorithms that automate the assignment of ICD-9-CM codes to anonymised radiology reports with a training set of 978 documents and a test set of 976 documents.

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).

In 2007 Computational Medicine Challenge [5] the top-performing systems achieved F-measure 0.8908, the minimum was with F-measure 0.1541, and the mean was 0.7670, with a standard deviation of 0.1340. Some 21 systems have F-measure between 0.81 and 0.90. Another 14 systems have F-measure 0.70. The article [9] compares three machine learning methods on radiological reports and points out that the best F-measure is 77%.

The top rated systems use variety of approaches like: Machine-learning; Symbolic methods; Hybrid approaches; UMLS Structures, Robust classification algorithm (naive Bayes) etc. the system reported in [8] uses a hybrid approach combining example-based classification and a simple but robust classification algorithm (naive Bayes) with high performance over 22 million PRs: F-measure 98,2%; for about 48% of the medical records at Mayo clinic. SynDiKATe [7] based on combination between text parsing and semantic information derivation from a Bayesian network and reports about 76% F-measure.

The better systems process negations, hypernyms and synonyms and were apparently doing significant amounts of symbolic processing.

SVMs and related approaches to machine learning were strongly represented in this challenge, but did not seem to be reliably predictive of high ranking. This motivated us to try to use SVM method for ICD-10 codes assignment to diagnoses, but enhanced with some preprocessing techniques applied to input data, concerning usage of synonyms, hyponyms, negation processing and word normalization and etc. used in other methods with better performance.

3 Material

The IE experiments were performed on training corpus if 1,300 and test corpus of 6200 anonymised hospital PRs for patients with endocrine and metabolic diseases provided by the University Specialised Hospital for Active Treatment of Endocrinology (USHATE), Medical University Sofia, Bulgaria.

Bulgarian medical texts contain a specific mixture of terminology in Latin, Bulgarian and Latin terms transcribed with Cyrillic letters (Table 1). There is no preferred language for the terminology so the two forms are used like synonyms. The terms occur in the text with a variety of wordforms which is typical for the highly-inflexional Bulgarian language.

The mixture of such terminology, given in Cyrillic and Latin alphabets, makes very hard the task for automatic assignment of ICD-10 codes to diagnoses. About 2.34% of the text is presented with Latin letters; the rest is written with Cyrillic symbols but contains Latin terminology (mostly diagnoses, anatomic organs and examinations) which is transliterated to Cyrillic alphabet (about 11.6% of all terms). About 37% of all diagnoses in our test corpus of 1,300 PRs were presented in Latin. This very specific medical language reflects the established medical tradition to use Latin language. Last but not least the foreign terminology is due to the lack of controlled vocabularies in Bulgarian

language. In addition no bilingual Bulgarian-Latin medical dictionary is available in electronic format as well.

Table 1 Examples for diagnoses representation

Type	Example
Mixture of medical terminology in Latin and Bulgarian	Консултация с офталмолог: ВОД= 0,6 ВОС=0,6, двуочно 0,8 с корекция. Фундоскопия: папили на нивото на ретината. Angiosclerosis vas. retinae hypertonica. Начални промени по типа на диабетна ретинопатия.
Medical terminology in Bulgarian	Диагноза: Захарен диабет тип 2. Затлъстяване II ст. Диабетна полиневропатия. Артериална хипертония-IIст. Диенцефален синдром.
Latin terms transcribed with Cyrillic letters	Диагноза: Хипопаратиреоидизмус постоператива компенсата. Хипотиреоидизмус Постоператива компенсата. Статус пост тиреоидектомиам про карцинома папиларе лоби синистри. Статус пост радиойодаблациам.

Further we have developed semi-automatically a dictionary with pairs of Latin and Bulgarian terms corresponding to anatomic organs and their status containing about 7,230 terms. The most complicated task was to develop semi-automatically the list of correspondences between diagnoses in Bulgarian and Latin. For this task were used resources available in Bulgarian [10] and English [11]: ICD-10 Classification (Fig. 1); Index of diseases and pathological states and their modifications (Fig. 2). There were also used: *(i)* Terminologia Anatomica providing terminology in English and Latin; *(ii)* Sets of about 300 prefixes and suffixes, about 100 roots, about 150 abbreviations in Latin and Greek and their corresponding meanings in English and Bulgarian; *(iii)* Rules for transliteration from Latin to Cyrillic.

E00.9	Вроден йод-недоимъчен синдром, неуточнен
E01.0	Дифузна (ендемична) гуша, свързана с йоден недоимък
E01.1	Полинодозна (ендемична) гуша, свързана с йоден недоимък
E01.2	Гуша (ендемична), свързана с йоден недоимък, неуточнена
E01.8	Други болести на щитовидната жлеза, свързани с йоден недоимък и сродни състояни
E03.0	Вроден хипотиреоидизъм с дифузна гуша
E03.1	Вроден хипотиреоидизъм без гуша
E03.2	Хипотиреоидизъм, дължащ се на лекарства и други екзогенни вещества
E03.3	Постинфекциозен хипотиреоидизъм
E03.4	Атрофия на щитовидната жлеза (придобита)
E03.5	Микседемна кома
E03.8	Други уточнени видове хипотиреоидизъм
E03.9	Хипотиреоидизъм, неуточнен
E04.0	Нетоксична дифузна гуша
E04.1	Нетоксичен единичен възел на щитовидната жлеза
E04.2	Нетоксична полинодозна гуша
E04.8	Други уточнени видове нетоксична гуша

Fig. 1 ICD-10 Classification in Bulgarian - excerpt for class "E"

Коронавирус (coronavirus), като причина за болест, класифицирана другаде B97.2
Коронарна (артерия) – *виж състояние*
Корсиканска треска (*виж също* Малария) B54
Кортико-адренален – *виж състояние*
Коса – *виж също състояние*
– изтръгване, скубане на F63.3
– – разстройство в стереотипа на движенията F98.4
Косвен – *виж състояние*
Косми
– враснали L73.1
– пръстеновидни или извити (вродени) Q84.1
Криза
– адисонова E27.2
– бъбречна N28.8
– вторична глаукома H40.4
– емоционална F43.2
– – остра реакция на стрес F43.0
– – приспособителна реакция F43.2
– – специфична за детството и юношеството F93.8
– коремна R10.4
– на Pel (табесна) A52.1
– надбъбречна (корова) E27.2
– нитритна
– – адекватно назначено и правилно приложено вещество I95.2
– – свръхдоза или погрешно вещество

Fig. 2 ICD-10 Volume 2 Tabular Index in Bulgarian – excerpt for "K" terms

ICD-10-CM (Clinical Modification) codes may consist of up to seven digits (Fig 3). A seventh character is required on some diagnoses that begin with "M," "O," "R," "S," "T," and "VWXY." and represents visit encounter or squeal for injuries and external causes.

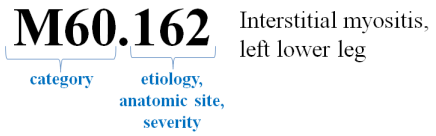

Fig. 3 ICD-10-CM Code Format

The ICD-10-CM is divided into the Alphabetic Index, an alphabetical list of terms and their corresponding code, and the Tabular List, a chronological list of codes divided into chapters based on body system or condition. The Alphabetic Index consists of the following parts: the Index of Diseases (Fig. 2) and Injury, the Index of External Causes of Injury, the Table of Neoplasms (Fig. 4) and the Table of Drugs and Chemicals [10].

	Злокачествено				С неопределен или неизвестен характер
	Първично	Вторично	In situ	Доброкачествено	
Новообразуване—*продължение*					
– външен(-на)					
– – маточна ос	C53.1	C79.8	D06.1	D26.0	D39.0
– – отвор (ухо)	C44.2	C79.2	D04.2	D23.2	D48.5
– вътрегръдна (кухина) (органи, НКД)	C76.1	C79.8		D36.7	D48.7
– вътреочно	C69.4	C79.4	D09.2	D31.4	D48.7
– вътрешна					
– – капсула	C71.0	C79.3		D33.0	D43.0
– – ос	C53.0	C79.8	D06.0	D26.0	D39.0
– вътрешни органи, НКД	C76.7	C79.8		D36.7	D48.7

Fig. 4 ICD-10 Volume 2 Table of Neoplasm

The data from ICD-10 Volume 2 Tabular index [10] are organized with leading term – level 1 (diagnose or pathological state) and modifications, which can be specified up to 7 levels. For instance for "A" terms are used 18256 words in total for explanation in different levels and 3568 different words.

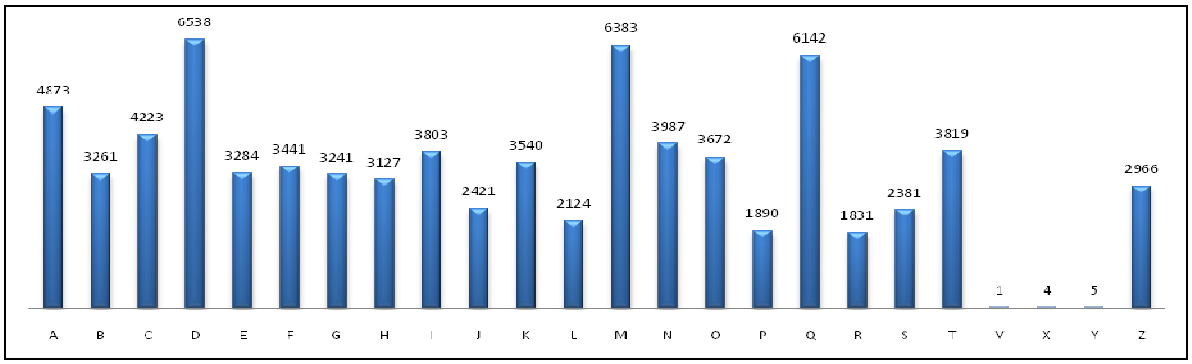

Fig. 5 Number of different diagnoses per cluster described in ICD-10 Volume 2 Tabular Index in Bulgarian

Searching in different sublevels not necessary specifies the ICD-10 codes. For instance, if the leading term is "Cyst" [10, 11], modification on level 1 "development" leads to K09.1, but further modification on level 2 "ovary" or "ovarian" leads to codes Q50.1 which belongs to other cluster. Another example for "Cyst" with modification on level 1 "epidermal" leads to L72.0 and further modification on level 2 "mouth" or "oral soft tissue" leads to codes K09.8. This shows that we need to use all nested levels of modifiers before final conclusion for the correct ICD-10 code for some diagnose.

In addition Tabular Index contains 19,161 different words and 291,116 words in total with repetitions, 2,221 in Latin (11.59%) and occurrences 83,713 in total (about 28.76%). This shows that direct application of Tabular index is not appropriate for automatic ICD-10 codes association for free-text diagnoses from discharge letters.

Tabular Index contains 76,939 descriptions of ICD-10 codes representing 9,044 different codes. ICD-10 classification [10] contains 14,439 different codes descriptions.

4 Method

4.1 SVM Classifier

SVMs have an ability to learn independent of the dimensionality of the feature space. This makes SVM Classifiers suitable approach for our task for automatic assignment of ICD-10 codes to diagnoses extracted from free-text PRs. This means that we can generalize even in the presence of very many features, because SVMs use overfitting protection, which does not necessarily depends on the number of features [12].

We present all ICD-10 codes as a set $C = \{c_1, c_2, \ldots, c_k\}$. The distribution in different clusters of codes described in the Tabular index (Fig. 5) shows that most of the diagnoses, except "VWXY" clusters, are described with variety of descriptions that in our opinion should be enough for generating rules for automatic classification.

To cover all possible codes included in ICD-10 classification we create training set of pairs (x_i, c_j) of diagnoses descriptions x_i and their corresponding ICD-10 codes $c_j \in C$ from extracted diagnoses descriptions from Tabular Index in Bulgarian, those used in ICD-10 classification and 1,300 PRs from training corpus. In the training set vectors x_i contains words used to describe diagnose with omitting meaningless word (e.g. a, an, the, this, that, and, or).

The implemented system (Fig. 6) works in two steps [13,14]: (i) Preprocessing and (ii) SVM Classification.

Preprocessing analysis includes several text processing tasks performed as pipeline: PRs sections splitting; Tokenization; Diagnoses extraction; Abbreviations expansion; Transliteration; Latin terminology processing; Words normalization; Medical terminology synonyms; SVM model.

Bulgarian hospitals discharge letters have mandatory structure [3]. The system splits the text on all available sections and passes *Diagnose* section text for further processing. *Diagnose* section text is spitted into words set $W = \{w_1, w_2, \ldots, w_p\}$. Using scoping rules applied to *Diagnose* section text words from the generated set W are combined into diagnoses $D = \{d_1, d_2, \ldots, d_n\}$. For each diagnose $d_m \in D$ we create vector $y_{mi} =< w_{m1}, w_{m2}, \ldots, w_{mq} >$ containing words included in it. Using sets AL and AB of abbreviations in Latin and Bulgarian language and functions $a_l : AL \rightarrow Lat$ and

$a_b : AB \rightarrow Bul$, words in vectors y_{mi} for each diagnose $d_m \in D$ are substituted by expanded terms meaning in Latin (Lat) and Bulgarian (Bul) language correspondingly according (1).

$$u_{mj} = \begin{cases} a_b(w_{mj}), if\ w_{mj} \in AB \\ a_l(w_{mj}), if\ w_{mj} \in AL \\ w_{mj}\ ,\ otherwise \end{cases} \quad (1)$$

Then we replace vectors y_{mi} by their corresponding vectors $z_{mi} =< u_{m1}, u_{m2}, \ldots, u_{mq} >$ for each diagnose $d_m \in D$. Using transliteration rules $t : Cyrillic \rightarrow Latin$ from Cyrillic to Latin alphabet we convert each word in vector z_{mi} to its equivalent in Latin. The cases when some words in vector z_{mi} are in Latin and the other are in Cyrillic are very rare. If $t(u_{mj}) \in Lat$ or $u_{mj} \in Lat$ we substitute it by its corresponding term $b_{mj} \in Bul$ from Bulgarian terminology repository *Bul,* otherwise we suppose that u_{mj} is in Bulgarian and set $b_{mj} = u_{mj}$. Using rules for words derivatives we replace all terms b_{mj} by their lemmas $l_{mj} \in Bul$ and construct vector $v_{mi} =< l_{m1}, l_{m2}, \ldots, l_{mr} >$. The result vector v_{mi} contains only words in Bulgarian. Further we process negations [15] and searching for synonyms and hyponyms of disease and pathological states names in Bulgarian medical terminology repository. We generate all possible partitions P_{mj} of consecutive words in vector v_{mi}:

$$P_{mj} = \{l_{m1}, \ldots, l_{mq} \mid l_{mq+1}, \ldots, l_{mt} \mid \ldots \mid l_{ms}, \ldots, l_{mr}\} \quad (2)$$

For each sequence in (2) in partition P_{mj} we create set of its synonyms s_l. Usually parts contain from 1 up to 7 consecutive words in the vector. The Cartesian product of synonym sets for partition P_{mj} generates set $Y_{mj} = s_1 \times \ldots \times s_q = \{y_1, \ldots, y_p\}$ of input vectors. The union (3) of these sets for all partitions contains input vectors with different descriptions of each diagnose d_m:

$$Y_m = \bigcup_{P_{mj}} Y_{mj} \quad (3)$$

We use the formal representation for SVM model, learned by training examples, to transform test examples as input vectors for SVM.

SVM Classification - The input space in SVM Classifier is a vector space and the output is a single number corresponding to different classes. **SVM** classifier applies binary classification for each of the input vectors $y_t \in Y_m$ for each diagnose $d_m \in D$ and each of the classes in C. Winning strategy ranks all classes and chooses the highest ranked class c_{im} for each diagnose $d_m \in D$.

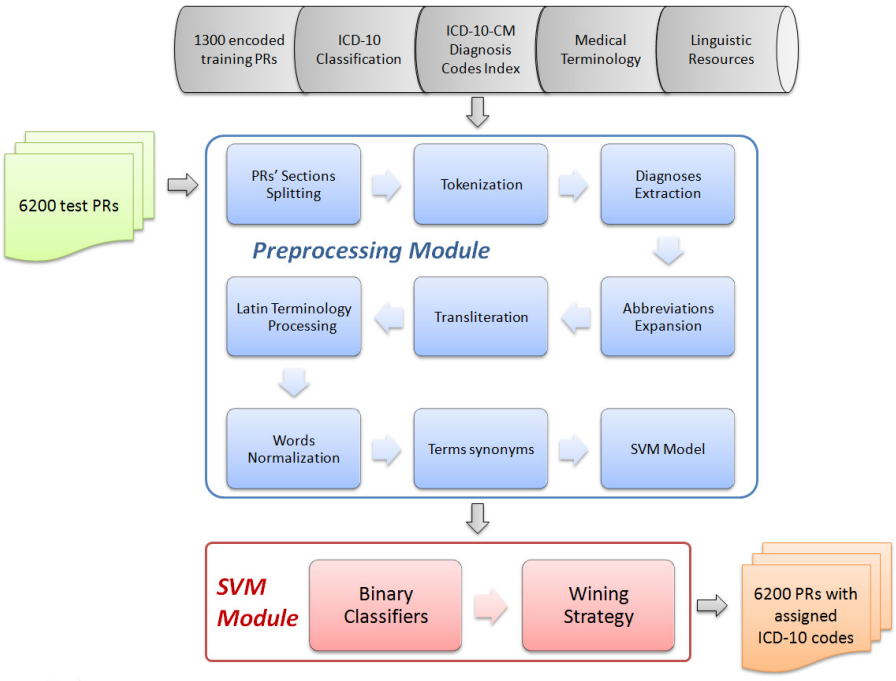

Fig. 6 System Architecture

4.2 Example

The implemented system allows processing of a single PR stored as text file in manual mode. There is also available automatic mode where can be processed all PRs stored in the selected folder and the result is stored in single CVS format file. In manual mode (Fig. 7) the text of PR is opened in section (1). After opening the text file, PR first is applied preprocessing steps from the algorithm and 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 (3) with recognized diagnoses within the text.

After selection of diagnose from list (3) to be processed its name is automatically excluded from list (3) and displayed in section (5). In the current example the selected diagnose *"киста оварии декстра" (киста на яйчника – in Bulgarian, cyst of ovary – in English)* is displayed in sections (5). The system identifies possible ICD-10 codes assignments and displays them in list (4) - *N83.0 Фоликуларна киста на яйчника (N83.0 Follicular cyst of ovary)*. 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. The data for processed diagnoses from list (3) are displayed in list (6) for further storage in CVS format text file.

In this example the diagnose *"феохромицитома" (pheochromocytoma)* is presented using Latin terminology with transliteration. This term corresponds to *"Доброкачествено новообразуване на надбъбречна жлеза" (neoplasm of Adrenal gland)* in Bulgarian language. In ICD-10 4 chars categories it corresponds to D35.0. The next diagnose *"киста оварии декстра" (киста на яйчника – in Bulgarian, cyst of ovary – in English)* is processed using again latin terminology transliteration for Latin term *"оварии" (ovarian) (яйчници – in Bulgarian)* and the result assigned code is *N83.0 Фоликуларна киста на яйчника (N83.0 Follicular cyst of ovary)*. Here *"декстра"* in Latin means (*дясна – in Bulgarian, Right – in English*) is not considered in classification in this case.

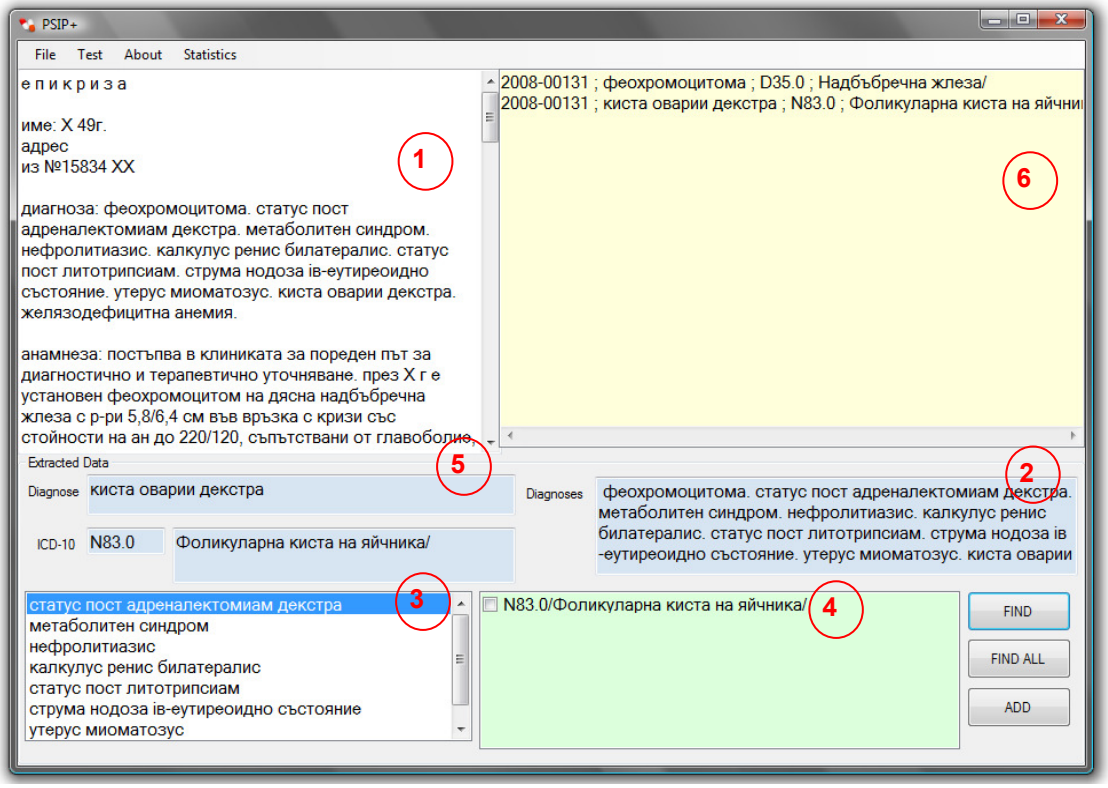

Fig. 7 Screenshot of System processing PR in manual mode

5 Evaluation and Results

The experiments were made with a training corpus described in Section 3 and the evaluation results are obtained using a test corpus, containing 6,200 PRs for patients with endocrine and metabolic diseases provided by USHATE.

For the test corpus there was identified descriptions of 26,826 diagnoses and 448 different classes diagnoses.

Because for the purposes of our project we are processing PRs for patients with endocrine and metabolic diseases their leading diagnoses are obviously classified in cluster "E". Thus some of the clusters are presented by few classified diagnoses (Fig. 8) and the other clusters (K, M, N, H, D, G, I) representing endocrine and metabolic diseases and related to them complications are presented by several classifications (Fig. 9).

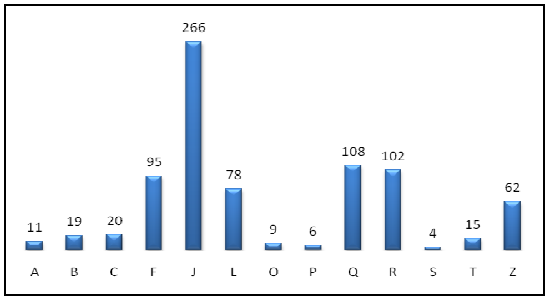

Fig. 8 Number of diagnoses classified for rare clusters

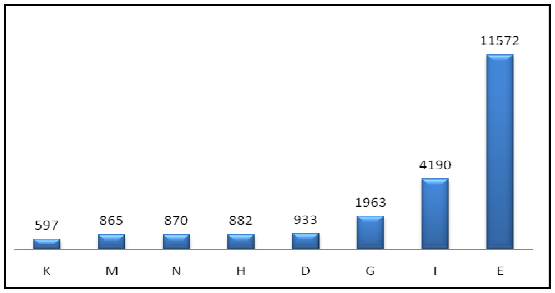

Fig. 9 Number of diagnoses classified for most common clusters

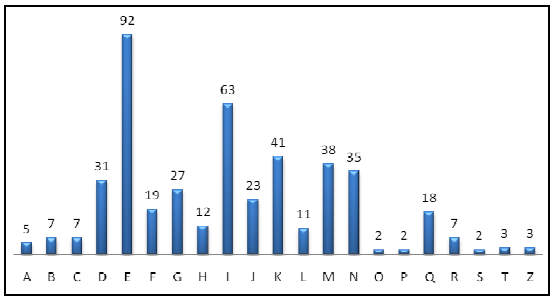

Fig. 10 Number of different diagnoses per cluster

Although the experiments was performed for such specific test set the diversity of result classes of diagnoses in test set presented on Fig. 10 and the average number of classified diagnoses per cluster (Fig. 11) shows that almost all cluster were presented by sufficient amount of examples.

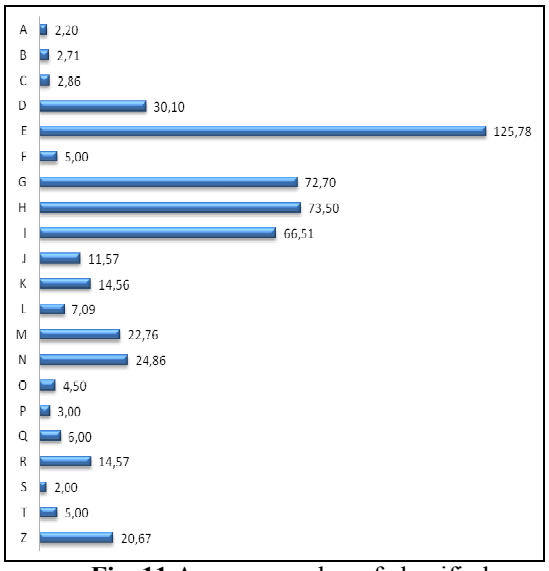

Fig. 11 Average number of classified diagnoses per cluster

Evaluation results (Table 3) shows high percentage of success in diagnoses recognition in PRs texts.

Table 3 Extraction sensitivity according to the IE performance measures

	Precision	**Recall**	**F-measure**
Diagnoses	97.3%	74.68%	84.5%

F-measure for leading diagnose in E-cluster and the diagnoses from the most common clusters (Fig. 9) is 98.76% for about 81.53% of the test set examples.

Obtained results are comparable with recent systems performing such task. For leading diagnoses we obtain better results, but still there are several difficulties like incorrect codes association due to:

- Latin terminology - for 345 cases;
- Abbreviations – for 538 cases
- Other – 1,202 cases describing mainly "status post" conditions, most of them is difficult to classify even manually.

For some diagnoses associated codes can be considered partially correct, because they the first three symbols of the ICD-10 code are assigned correctly, but the next tree symbols are either not specified, or associated to too general classes like *"... unspecified"*, *"... classified elsewhere"*, *"other disorders of ..."* etc.

6 Conclusion and Further Work

This paper presents software modules for ongoing scientific project which supports the automatic extraction of diagnoses from PR texts

The implemented modules are strictly oriented to Bulgarian language.

Usage of SVM method for ICD-10 codes assignment to diagnoses, enhanced with some preprocessing techniques applied to input data, concerning usage of synonyms, hyponyms, negation processing, word normalization, Latin terminology and abbreviations processing and etc. shows better performance in certain context.

The plans for their further development and application are connected primarily to Bulgarian local context. For diagnoses recognition task we plan improvement of rules and extension of resource bank for Latin terminology and abbreviations for more precise code assignments.

Acknowledgments

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).

References

[1] International classification od Diseases, World Health Organization, http://www.who.int/classifications/icd/en/

[2] 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), pp. 760-772.

[3] 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,Bulgaria*, http://dv.parliament.bg/

[4] 2007 International Challenge: Classifying Clinical Free Text Using Natural Language Processing, http://computationalmedicine.org/challenge/previous

[5] Pestian J, C. Brew, P. Matykiewicz, DJ Hovermale, N. Johnson, K. B. Cohen, and D. Wlodzislaw. A shared task involving multi-label classification of clinical free text. *In: ACL'07 workshop on biological, translational, and clinical language processing (BioNLP'07).* Prague, Czech Republic; (2007), pp. 36–40.

[6] 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.

[7] Hahn, U., M. Romacker and S. Schultz. Creating Knowledge Repositories from Biomedical Reports: The MEDSYNDIKATE Text Mining System, *In Pacific Symposium on Biocomputing,* vol. 7, (2002), pp. 338-349.

[8] 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.

[9] 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

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

[11] 2011 ICD-10-CM Diagnosis Codes Index, http://www.icd10data.com/ and http://www.cdc.gov/nchs/icd/icd10cm.htm#10update

[12] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In Claire N´edellec and C´eline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, Chemnitz, DE, Springer Verlag, Heidelberg, DE, (1998), pp. 137–142.

[13] 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, (2010), pp. 56-66.

[14] Tcharaktchiev, D., G. Angelova, S. Boytcheva, Z. Angelov, and S. Zacharieva. Completion of Structured Patient Descriptions by Semantic Mining. In Koutkias V. et al. (Eds), *Patient Safety Informatics, Stud. Health Technol. Inform.* 2011 Vol. 166, IOS Press, (2011), pp. 260-269.

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