Applying Language Technologies on Healthcare Patient Records for Better Treatment of Bulgarian Diabetic Patients

Ivelina Nikolova[1], Dimitar Tcharaktchiev[2], Svetla Boytcheva[3], Zhivko Angelov[4] and Galia Angelova[1]

[1] Institute of Information and Communication Technologies, Bulgarian Academy of Sciences (IICT-BAS), Sofia, Bulgaria

[2] Medical University – Sofia, Bulgaria

[3] American University in Bulgaria, Blagoevgrad, Bulgaria

[4] Adiss Lab Ltd. Sofia, Bulgaria

iva@lml.bas.bg, dimitardt@gmail.com, sboytcheva@aubg.bg, angelov@adiss-bg.com, galia@lml.bas.bg

Abstract. This paper presents a research project integrating language technologies and a business intelligence tool that help to discover new knowledge in a very large repository of patient records in Bulgarian language. The ultimate project objective is to accelerate the construction of the Register of diabetic patients in Bulgaria. All the information needed for the Register is available in the outpatient records, collected by the Bulgarian National Health Insurance Fund. We extract automatically from the records' free text essential entities related to the drug treatment such as drug names, dosages, modes of admission, frequency and treatment duration with precision 95.2%; we classify the records according to the hypothesis “having diabetes” with precision 91.5% and deliver these findings to decision makers in order to improve the public health policy and the management of Bulgarian healthcare system. The experiments are run on the records of about 436,000 diabetic patients.

Keywords: Biomedical Natural Language Processing, Business Intelligence, Big Data

1 Introduction

The constant growth of electronic narratives discussing patient-related information implies constant growth of the attempts to process these texts automatically. It is well known that the most important findings about the patients are kept as free texts in various documents and languages but these text descriptions are usually oriented to human readers. In this way Information Extraction (IE) becomes the dominating natural language processing (NLP) approach to biomedical texts. The main IE idea is to extract automatically important entities from free texts, with accuracy as high as

possible, and to build software systems operating on these entities (skipping the non-analysed text fragments). NLP in general is viewed as a rather complex Artificial Intelligence task so IE is proposed as a technology at the middle between keyword search and deep text analysis; it focuses on surface linguistic phenomena that can be recognised without deep inference. The NLP performance gradually improves during the last decades but the IE systems are still rarely used outside the research groups where they have been developed [1]. On the other hand it is expected that IE progress would enable radical improvements in the clinical decision support, biomedical research and the healthcare in general [2]. Leading industrial companies claim that NLP is an enabling technology and can be leveraged today for revenue, efficiency and quality but some 10-15 years are needed for mature technological development [3].

Recent Big Data challenges add a new perspective to the complex task of secondary use of electronic health records. Today typical collections of clinical narratives contain millions of records for millions of patients so even procedures for pseudonymisation and anonymisation are problematic. In this paper we present a project dealing with dozens of millions of outpatient records where NLP is carefully applied to specific text sections of the patient records. The extraction components in use are developed several years ago, continuously upgraded, tested and evaluated to deliver entities extracted with high accuracy. We present an integrated system of a business intelligence tool and NLP modules performing big data normalisation, cleaning and knowledge discovery and show a use case on medical data. We demonstrate that the combination of these tools can help to build the Register of diabetic patients by discovering potential diabetic patients which were not formally diagnosed with diabetes. The IE focus is on the patients' medical treatment and patient anamnesis.

Section 2 presents the project objectives and the data repository we use. Section 3 describes the Business Intelligence tool *BITool*. Section 4 presents the knowledge discovery modules, their application to real data and evaluates their performance. Section 5 sketches further work and the conclusion.

2 Building the Bulgarian Diabetic Register

The ultimate project objective is to accelerate the construction of the Register of diabetic patients in Bulgaria by integration of language technologies and business intelligence tools. Advanced information technologies would enable to: *(i)* keep the established practice of patient registration without burdening the medical experts with additional paper work; *(ii)* reuse the existing standard records in compliance with all legal requirements for safety and data protection; *(iii)* save time and resources by avoiding multiple patient registrations and disturbance of the diagnostic and treatment process. Practically, once entered in the healthcare system, the patient data might be reused in multiple aspects. Multiple registrations and growing administrative burdens are seen as a major obstacle for the development of the Register. A web-interface for self-registration to the Bulgarian Diabetic Register is foreseen as well.

The Register contains 28 indicators of diabetic patients' status, including age, sex, ICD-10 codes of diagnoses of diabetes and its complications, diabetes duration, risk

factors, data about compensation, laboratory results, hospitalisations and prescribed medication. Manual collection of data proved to be impractical during the last ten years; in addition there are many diabetic patients who are not formally diagnosed and not treated at all. In the case of diabetes, a progressive chronic disease with serious complications, it is highly desirable to develop a system for early alerts that might signal eventual diabetes symptoms.

It turns out that all the information needed for the Register is available in the outpatient records, collected by the Bulgarian National Health Insurance Fund. There are multiple records stored for the same patient along the months and the years. Given that this information is extracted automatically, a Business Intelligence tool can deliver various types of findings to decision makers in order to improve the public health policy and the management of Bulgarian healthcare system. Actually the *BITool* is useful anyway because the data of the Health Insurance Fund contains a lot of information that is structured using codes of medical classifications and nomenclatures. However in this paper we are interested in the analysis of free texts and capturing some essential entities described there. By means of NLP techniques integrated with the *BITool* we discover the potential diabetic patients which were not formally diagnosed with diabetes.

Thanks to the support of the Bulgarian Ministry of Health and the National Health Insurance Fund, the Medical University - Sofia has received for research purposes a large collection of outpatient records. The data repository currently contains more than 37.9 million pseudonymised reimbursement requests (***outpatient records***) submitted to the National Health Insurance Fund (NHIF) in 2013 for more than 5 million patients, including 436,000 diabetic ones. In Bulgaria the outpatient records are produced by the General Practitioners (GPs) and the Specialists from Ambulatory Care for every contact with the patient (in patient home as well as in doctors' office).

The outpatient records are semi-structured files with predefined XML-format. Despite their primary accounting purpose they contain sufficient text explanations to summarise the case and to motivate the requested reimbursement. The most important indicators like *Age*, *Gender*, *Location*, *Diagnoses* are easily seen since they are stored with explicit tags. The Case history is presented quite briefly in the *Anamnesis* as free text with description of previous treatments, including drugs taken by the patient beyond the ones that are to be reimbursed by the Insurance Fund. *Family history* and *Risk factors* are often included in the *Anamnesis* of diabetic patients. *Patient status* is another section containing free text. It includes a summary of the patient state, symptoms, syndromes, patients' height and weight, body mass index, blood pressure and other clinical descriptions. The values of *Clinical tests and lab data* are enumerated in arbitrary order as free text in another section. A special section is dedicated to the *Prescribed treatment*. Only the drugs prescribed by the GPs and reimbursed by the NHIF are coded, using the specific NHIF nomenclatures. All the other medications and treatment procedures are described as free text. In contrast to clinical discharge letters that might discuss treatments in longer past and future periods, the *Prescribed treatment* section in the outpatient records is more focused to the context at the moment when the record is composed.

The repository given to the Medical University – Sofia is pseudonymised by NHIF which has the keys for mapping the records to the original patients. Our experiments use a completely anonymised data set. Fortunately, using the pseudonymised patient identifier, it is possible to track automatically the multiple visits of the same patient to GPs and Ambulatory Care, which is important in the case of a chronicle disease like diabetes.

An outpatient record might include about 160 tags. The average length of the files is about 1 MB. For our purposes, we work with about 20-30 tags and consider the unstructured content of four sections.

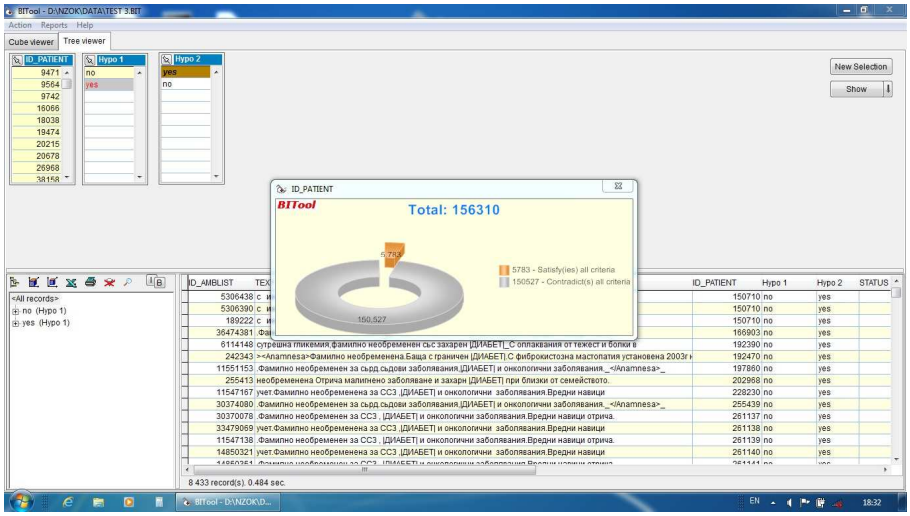

Figure 1. A *BITool*-constructed concordancer for the word DIABETES in the *Anamnesis* of outpatient records for patients who are not formally diagnosed with Diabetes Mellitus (E11)

3 BITool and Data Integration

BITool serves as an integration platform of the analyses performed on the medical data. It offers means to operate on the original data, such as search facilities and means to extract subsets of semi-structured data, as well as to enrich the stored resources with newly extracted features (e.g. from the NLP or statistical modules).

BITool offers powerful functionalities for online analytic processing (OLAP) of large data repositories, organised as a *n*-dimensional cube. The cube is easy to construct by drag-and-drop interface after selection of the desired attributes and their values (see 3 selected dimensions of Fig. 1 upper left). Fig. 1 shows the result of an intermediate step of the task to investigate whether some patients without formal diagnosis of diabetes (code E11) have symptoms typical for this disease. This can be discovered after finding text descriptions of typical events: e.g. mentioning the phrase "*diabetic foot*" in a positive context. Concordancers of text fragments using a 12-words window (6 words before and 6 words after) are constructed around the words

of interest. In Fig. 1 the records of 156,310 patients, who are not formally diagnosed with diabetes mellitus, are checked for occurrences of the word "*diabetes*" in the *Anamnesis* section. The records of 5,783 patients contain the word "*diabetes*". These are later used in the knowledge discovery phase by the NLP modules. After the processing the results are re-integrated back to *BITool* to be used in further analyses and research or health management tasks such as to help the construction of the diabetic register.

4 Knowledge Discovery

Our goal in this study is to recognise patients who have diabetes but have not been formally diagnosed with this disease. As hints signalling the presence of diabetes in the outpatient record we consider (*i*) the medical treatment – if the patient has diabetes he/she would also take appropriate drugs, and (*ii*) statements in the anamnesis about the patient having diabetes or its complications. We analyse the respective outpatient record sections by applying NLP over feature vectors extracted by the *BITool*.

IE from free texts finds entities of interest by focusing on important words and phrases that trigger shallow analysis in the local context. The texts in the outpatient records are written in a specific medical sublanguage containing mostly phrasal structures, terms in Bulgarian and Latin, typical abbreviations etc. Grouping together the records for the same patient we can track the progress of diabetes and its complications. The implemented IE components extract with high accuracy information about patient status, current treatment, hospitalisation, diabetes compensation, family history and risk factors, as well as values of specific clinical tests and lab data. Identifying values of lab tests is important since e.g. blood sugar levels are a typical signal of diabetes. This extractor uses a rule-based approach for recognition of linguistic patterns corresponding to the entities of interest. Major difficulties encountered in the development are due to the large variety of expressions describing the laboratory tests and clinical examinations. Here we consider in more details the drug extractor [4] that has been extended recently to cope with the NHIF outpatient records.

4.1 Structuring drug treatment information

An automatic procedure analyses the free texts in the *Prescribed treatment* section in order to extract information about: drug names; dosages; modes of admission; frequency and treatment duration. It assigns the corresponding ATC[1] and NHIF codes to each medication event. The list of registered drugs in Bulgaria is provided by the Bulgarian Drug Agency; it contains about 4,000 drug names and their ATC codes.

The extraction is based on algorithms using regular expressions to describe linguistic patterns. There are more than 80 different patterns for matching text units

1 Anatomical Therapeutic Chemical (ATC) Classification System for the classification of drugs, see http://www.who.int/classifications/atcddd/en/

which deal with the ATC and NHIF code, medication name, dosage and frequency. Some regular expressions are illustrated at Fig. 2.

For diabetic patients, currently the extractor handles 2,239 drugs names included in the NHIF nomenclatures. Recent extraction evaluation has been performed with large-scale analysis of the outpatient records of 33,641 diabetic patients for 2013. The precision is 95.2% and the sensitivity - 93.7%. This result is slightly better than the accuracy reported in 2011 [5] when the extractor was a (research) prototype dealing with less than 500 drugs. The performance of the module is evaluated manually. The labelled data is split to 20 equal subsets and randomly selected records are evaluated by an expert (about 40% of each subset). The average of the subset evaluation is the final score of the module.

Figure 2. Regular expressions of linguistic patterns for analysis of Dosage

The major reasons for incorrect recognition of drug events are: *(i)* misspelling of drug names; *(ii)* drug names occurring in the contexts of other descriptions; *(iii)* undetected descriptions of drug allergies, sensibility, intolerance and side effects; *(iv)* drug treatment described by (exclusive) *OR*; *(v)* negations and temporally interconnected events of various kinds: undetected descriptions of cancelled medication events; of changes or replacements in therapy; of insufficient treatment effect and change of therapy.

About 30% of the medication events in the test corpus were described without any dosage. Lack of explicit descriptions occurs mostly for treatment of accompanying diseases. After applying the recognition algorithm and default daily dosage, the number of records lacking dosage has been reduced to 15.7% in the final result.

4.2 Discovering potential diabetic patients

Now we consider in more detail the discovery of potential diabetic patients that are not formally diagnosed in the NHIF repository for 2013. Medical experts propose criteria for happening of the event "*having diabetes*": e.g. high blood sugar or high glycated hemoglobin in the text of the section *Lab test results*, or admission of drugs used for diabetes treatment mentioned in the *Anamnesis*, or statements in the *Anamnesis* describing diabetes or its symptoms. A concordancer is built on the outpatient records for words related to descriptions of such events. The focused records are subject to further NLP analysis in order to confirm the relevant hypotheses. Iteratively the set of records under consideration can be shrunk by excluding the already checked patients. The *BITool* checks a single hypothesis for about 40 seconds. Checking several hypotheses simultaneously takes approximately the same time for each new one. OLAP is used for checking hypotheses and

calculating the support to corresponding associative rules. The results of text processing at records level are re-integrated back to *BITool* as additional attributes to each outpatient record. These results can be manual annotations (considered "trusted") or automatically generated labels (called "generated"). The process is illustrated at Fig. 3.

4.2.1. Filtering diabetes related records

To confirm/reject the hypothesis of *having diabetes*, we shall apply supervised machine learning (ML) methods for filtering text chunks from the free text zones of the outpatient record. The focus of this example is the word *diabetes* but the same analysis can be performed for any other disease, symptom, word or phrase of interest. The selected outpatient records have no explicit diagnosis of diabetes. We want to show that NLP can help to find contradiction in the repository (thus supporting data cleaning) and/or to extend the records' metadata when matching free text fragments confirming the diabetes diagnosis.

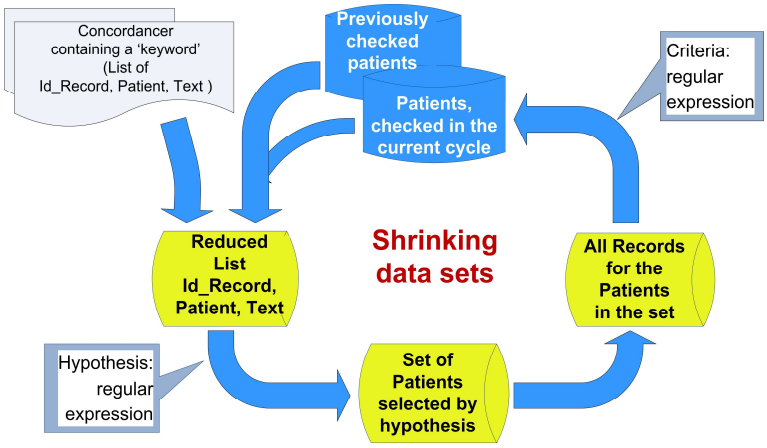

Figure 3. Iterative shrinking of sets of outpatient records to confirm/reject a hypothesis

4.2.2. Input data

The input data are text chunks extracted from a concordancer built for the string *диабет* (*diabetes*). The whole data set consists of 67,904 distinct chunks extracted from the records of 156,310 patients who are not formally diagnosed with diabetes. Each chunk contains the word *diabetes* and a 6-token window of its left and right context. The text is only tokenised and stemmed. Our procedure will perform a classification task - confirm/reject the hypothesis without recognising the particular trigger expressions.

Here follow some sample chunks which demonstrate the variety of positive and negative examples:

(i) NEG Фамилност- обременен/а-**диабет**ици по майчина линия/.

*Family - heredity - **diabet**ic on maternal line.*

(ii) NEG Необходимо е изключване на стероиден **диабет**; насочва се към ТЕЛК...

*It is necessary to exclude steroid **diabet**es; re-directing to TEMC...*

(iii) POS Покачва кръвно налягане. Има **диабет**. Оплаква се от сърцебиене...

*Raises the blood pressure. Has **diabet**es. Complains of palpitation...*

Examples (*i*) and (*ii*) are negative - the first one shows only heredity but does not confirm the diagnosis *diabetes*, in the second one the diagnosis is negated. However example (*iii*) is positive. Often the positive diabetes statement is given within 3-4 tokens.

4.2.3. Experiments and results

We apply a combined rule-based and machine-learning approach to solve the problem: (*i*) iterative rule-based rough filtering to reduce the size of the data that potentially confirms our hypothesis; (*ii*) supervised classification of the records in the reduced dataset to train a model which classifies the records as positive or negative to the hypothesis.

Rule-based rough pre-filtering

The data is extracted from records where *diabetes* is not explicitly mentioned among the diagnoses; therefore, we expect that most of the input chunks are negative examples. We decided first to filter as many of the negative ones as possible in a rule-based approach and reduce the size of the input that will be classified further. Several patterns of the negative examples appeared quite often in our data - e.g. "*no evidence about diabetes*", "*no diabetes in the family*" etc., which match about 10% of the input records. These are expressions talking about family heredity or rejection of the diabetes diagnosis. Iteratively we created a set of regular expressions on stem level which match such chunks. With a set of 41 expressions in the filter, the number of chunks was reduced to 26,000 (which is about 1/3 of the initial corpus size).

Supervised classification of positive/negative examples

In the second phase we extracted from the reduced data set two random subsets – one of 282 documents and one of 1,000 documents and we annotated them. The first one was a development set, we used it for selecting the features and make initial tests. It contains 74 positive and 208 negative examples whereas the second one contains 187 positive and 813 negative examples. By using various features and classification algorithms we check the applicability of ML to the automatic extraction of records referring to "*having diabetes*" (similarly to [6, 7, 8]) and set a reasonable baseline for this task. The difference in our approach is that we model and process real world big data.

In the preprocessing phase we stem the chunks to overcome the morphological inflection (Bulgarian is highly inflectional). Due to the large number of abbreviations, dosages, lab test results etc. which contain punctuation marks, the automatic sentence splitting with the available tools is unreliable and we do not apply it.

By achieving comparatively high results with surface and structural features only we prove the applicability of the approach. We experimented with both boolean and nominal feature vectors. The vectors characterising each chunk in the classification task correspond to word stems, bigrams and trigrams which are pre-defined. In the boolean settings the feature is true if the corresponding attribute - stem/bigram/trigram is available in the text chunk and negative otherwise. The list of stem-features is constructed by analysing the set of stems which occur in the datasets. We made experiments with 3 different feature corresponding to the experiments described below.

We tried several algorithms on the same dataset: NaiveBayes, J48, SMO and JRip, all with boolean features: JRip and J48 performed best. We did also classification with nominal features with MaxEnt algorithm and this one outperformed all the rest in means of precision. The features we used were the same - the words' stems, bigrams and trigrams.

In this study we are most interested in the precision on positive examples because in this way we can add to the register (with minimal manual effort/check) with high certainty the correctly recognised positive examples as patients *having diabetes* who were not formally diagnosed with diabetes. By applying several high precision filters like one ones scatched above we could achieve also also good coverage in the end. Here follow details about the experiments. The results are available in Tables 1, 2 and 3.

Experiment 1

We use 93 features which correspond to stems of terms occurring in the text of positive examples (excluding numbers). The results from a 10-fold cross-validation with the best algorithms J48 and JRip are shown in the first column of Table 1. J48 recognised only 63.1 % of the positive examples however the precision achieved by JRip is encouraging – 91.2%. The rules inferred by JRip are only 2 but obviously they fit well the data.

Rule 1*:* (*family*) = *false*) and (*sugar* = *true*) and (*noninsulindependent*= *true*) =>
class=pos (30.0/2.0)

Rule 2*:* (*family* = false) and (*sugar* = true) and (*treat* = true) =>
class=pos (6.0/0.0)

However on a larger scale these two rules might not suffice for achieving such good performance and we kept elaborating our features.

Experiment 2

We use 112 textual features which correspond to the stems of terms occurring in positive and negative examples. The terms are pre-filtered manually by an expert.

Both algorithms - JRip and J48 performed worse than in the first experiment and scored precision under 70% (Table 1, Exp.2).

Experiment 3

In this experiment we took advantage of the automatic feature-selection algorithms. Our initial feature set contained 10,576 attributes corresponding to the stems of all terms in the development dataset. We applied on it the chi-squared attribute evaluator implemented in Weka [9] according to the feature representativeness in the development set. As result 151 features were selected. In Experiments 3a and 3b we used them in combination with bigrams and with bigrams and trigrams respectively (Table 2). The results of both algorithms improve by adding bigrams and by including trigrams rise even more. JRip reaches 65.3% precision and 73.1% when adding trigram features. J48 achieves precision as high as 86.1% after adding bigrams and 87.2 including trigrams in the feature set. In the tree built by J48 one could clearly see the importance of bigrams and trigrams features - out of 17 tree leaves, only 4 are unigrams; the rest are bigrams and trigrams. We explain this with the fact that trigrams capture the order of the tokens and represent concrete expressions signalling the presense/absense of diabetes such as: "*инсулинозависим захарарен диабет*" (*insulindependent diabetes mellitus*), "*неинсулинозависим захарен диабет" (noninsulindependent diabetes mellitus), "със зах диабет" (with diabetes mellitus), "майка с диабет*" (*mother with diabetes*). According to our observations the last feature signals the absence of diabetes in the concrete excerpt because normally in the length of one excerpt, if there is a family anamnesis, there is no other description related to the patient which signals diabetes (the length is too small to fit more descriptions along with a family anamnesis). The results suggest that until this point adding more features results in better precision (except for JRip in Experiment 1) and also growing recall when using JRip.

	Exp. 1: 93 pos features; stems only; 10-fold cross-validation						**Exp. 2: 112 pos/neg features; stems only; 10-fold cross-validation**					
	JRip			**J48**			**JRip**			**J48**		
Class	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
positive	**91.2**	16.6	28.1	66.7	21.4	32.4	61.1	29.4	39.7	**65.8**	52.4	58.3
negative	83.9	99.6	91.1	84.4	97.5	90.5	85.5	95.7	90.3	89.5	93.7	91.6
weighted avg	85.2	84.1	84.1	81.1	83.3	79.6	80.9	83.3	80.8	85.1	86	85.4

Table 1. Results from Experiment 1 and 2 with manually selected features

	Exp. 3a: 151 automatically selected stem features + bigrams; 10-fold cross-validation						Exp. 3b: 151 automatically selected stem features + bigrams + trigrams; 10-fold cross-validation					
	JRip			J48			JRip			J48		
Class	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
positive	65.3	41.2	50.5	**86.1**	36.4	51.1	73.1	40.6	52.2	**87.2**	36.4	51.3
negative	87.5	95	91.1	87.1	98.6	92.5	87.6	96.6	91.9	87.1	98.8	92.6
weighted avg	83.4	84.9	83.5	86.9	87	84.8	84.9	86.1	84.5	87.1	87.1	84.9

Table 2. Experiments 3a and 3b - automatic feature selection, bigrams and trigrams.

	MaxEnt					
	Exp. 4a: all stems + bigrams; 10-fold cross validation			Exp. 4b: all stems+bigrams + trigrams; 10-fold cross validation		
Class	**P**	**R**	**F**	**P**	**R**	**F**
positive	91.3	22.6	36.2	**91.5**	20	32.8
negative	85.6	84.2	85	85.6	84.2	84.9
weighted avg	88.45	53.4	60.6	88.55	52.1	58.9

Table 3. Experiments 4a, 4b – MaxEnt with all textual features, bigrams and trigrams.

Experiment 4

We trained a model with MaxEnt algorithm using all textual features plus bigrams (Exp. 4a) and bigrams and trigrams (Exp. 4b) as nominal values. All strings were first stemmed. These two experiments gave almost the same results and outperformed J48 and JRip in means of precision. The best precision on positive examples we reached is as high as 91.5% when including bigrams and trigrams. Using MaxEnt with nominal features has the advantage that the features are not pre-set. The similar results also suggest that using MaxEnt with these features could be set as a reasonable baseline.

The positive examples in our database are so rare (because in principle diabetic patients *are* formally diagnosed) that the data quantity has major impact on the training. Having a larger corpus (which could happen when records from previous years become available) and having more golden data will help for learning better the patterns of positive examples. Nevertheless the current results show that such a hybrid method combining rule-based and machine learning approach can be used to prove the hypothesis *having diabetes* with high precision.

5 Conclusion and Further Work

The paper presents our first steps towards building a computational platform for tackling Big Data in the medical domain in a real-world application. It is obvious that decisions about medical cases cannot be made completely automatically so the final judgment should be always subject to human considerations. But the automatic processing of texts in the patient records defines a completely new horizon for most tasks related to health analytics.

The IE modules are exploited quite carefully, for extraction of a limited number of entities and events only. They are tested in various scenarios and gradually improve their performance using hybrid rule-based and machine learning approaches. We extract automatically from the records' free text essential entities related to the drug treatment such as drug names, dosages, modes of admission, frequency and treatment duration with precision 95.2%; we classify the records according to the hypothesis "*having diabetes*" with precision 91.5% and deliver these findings to decision makers

in order to improve the public health policy and the management of Bulgarian healthcare system. We think that large-scale analysis of medical texts can be viewed as a reliable technology if the input is well-structured into zones (which is the case of the outpatient records) and the extraction task has clear and well-defined target entities.

6 Acknowledgements

The research work presented in this paper is partially supported by the FP7 grant AComIn No. 316087, funded by the European Commission in the FP7 Capacity Programme in 2012–2016. The team acknowledges also the support of Medical University – Sofia, the Bulgarian Ministry of Health and the Bulgarian National Health Insurance Fund.

7 References

1. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.F.: Extracting Information from Textual Documents in the EHR: A Review of Recent Research. In: Geissbuhler, A., Kulikowski, C. (eds.) *IMIA Yearbook of Medical Informatics*, pp. 138–154 (2008).
2. Demner-Fushman, D., Chapman, W., McDonald, C.: What can NLP do for Clinical Decision Support? *J. of Biomedical Informatics* 42(5), 760–772 (2009).
3. Health Fidelity: The What, When, Where and How of Natural Language Processing. *NLP issue brief*, 2013, http://healthfidelity.com/technology/issue-briefs/nlp-issue-brief
4. Boytcheva, S. Shallow Medication Extraction from Hospital Patient Records. In: Koutkias, V., J. Niès, S. Jensen, N. Maglaveras, and R. Beuscart (Eds.), *Studies in Health Technology and Informatics series*, Vol. 166, IOS Press, 119-128 (2011).
5. Boytcheva, S., D. Tcharaktchiev and G. Angelova. Contextualization in automatic extraction of drugs from Hospital Patient Records. In A. Moen at al. (Eds) User Centred Networked Health Case, IOS Press, *Studies in Health Technology and Informatics*, Vol. 169, pp. 527-531 (2011).
6. Nikolova, I. Unified Extraction of Health Condition Descriptions, *Proceedings of the NAACL HLT 2012 Student Research Workshop*, June 2012, Montreal, Canada, pp. 23—28. Available at http://aclweb.org/anthology//N/N12/N12-2005.pdf.
7. Savova G., P. Ogren, P. Duffy, J. Buntrock, and C. Chute. Mayo Clinic NLP System for Patient Smoking Status Identification. *Journal of American Medical Informatics Association*, 15(1), pp. 25-28 (2008).
8. Chapman D. Chu, J.N. Dowling and W.W. Chapman. Evaluating the Effectiveness of Four ContextualFeatures in Classifying Annotated Clinical Conditions in Emergency Department Reports. *Proceedings of AMIA Annual Symposium* 2006, pp. 141-145.
9. Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/