Extraction and Exploration of Correlations in Patient Status Data

Svetla Boytcheva[1], Ivelina Nikolova[2], Elena Paskaleva[2], Galia Angelova[2],
Dimitar Tcharaktchiev[3] and Nadya Dimitrova[4]

[1] *State University of Library Studies and Information Technologies, Sofia, Bulgaria, svetla.boytcheva@gmail.com*
[2] *Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria, {iva, hellen, galia}@lml.bas.bg*
[3] *University Specialized Hospital for Active Treatment of Endocrinology, Medical University, Sofia, Bulgaria, dimitardt@gmail.com*
[4] *National Oncological Hospital, Sofia, Bulgaria, dimitrova.nadia@gmail.com*

Abstract

The paper discusses an Information Extraction approach, which is applied for the automatic processing of hospital Patient Records (PRs) in Bulgarian language. The main task reported here is retrieval of status descriptions related to anatomical organs. Due to the specific telegraphic PR style, the approach is focused on shallow analysis. Missing text descriptions and default values are another obstacle. To overcome it, we propose an algorithm for exploring the correlations between patient status data and the corresponding diagnosis. Rules for interdependencies of the patient status data are generated by clustering according to chosen metrics. In this way it becomes possible to fill in status templates for each patient when explicit descriptions are unavailable in the text. The article summarises evaluation results which concern the performance of the current IE prototype.

Keywords

Medical Information Extraction, Template Filling, Correlations of Patient Status Data

1. Introduction

Patient data are stored in various formats including paper archives which are recently transformed to electronic files. The task of Information Extraction (IE) from patient records is very important, because it enables automatic generation of databases with structured patient data that can be explored for improving the diagnostics, care decisions, the personalised treatment of diseases as well as other fields like the healthcare management, epidemiology etc. On the other hand most of the clinical documents present only partial information about the patients so some kind of aggregation is needed to provide a complex view to the patient health status.

Medical documents contain much unstructured text and their automatic processing is a challenge to be faced for every natural language separately. Especially for Bulgarian language, the biomedical NLP is making its initial steps. This is partly due to the lack of large corpora in the medical domain. Our present work deals with anonymous hospital records of patients who are diagnosed with different forms of diabetes. The PR pseudonymisation is done by the information system of the University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev", which is part of the Medical University – Sofia. The general objective of our project, to be achieved by 2011, is to apply IE techniques to patient texts in order to extract data concerning the hospitalisation effect. Currently we consider the extraction of patient status data, i.e. the recognition and structuring of the patient's symptoms which are related to diabetes.

Bulgarian medical texts contain a specific mixture of terminology in Latin, Cyrillic and Latin terms transcribed with Cyrillic letters. The terms occur in the text with a variety of wordforms which is typical for the highly-inflexional Bulgarian language. The major part of the text consists of sentence phrases without agreement and often without proper punctuation marks. We consider normalised texts with standard abbreviations and without spelling errors, because we aim at a research study. Our present experimental corpus is formed by some 6400 words, with some 2000 of them being medical terms.

The paper is structured as follows: Section 2 overviews some related research and discusses IE applications in the medical domain. Section 3 describes the raw data features, the main types of patient status data and some techniques for their extraction. Section 4 presents the approach for aggregation of patient data and calculation of correlations among different values of patient characteristics. Section 5 sketches the evaluation of the present IE prototype. Section 6 contains the conclusion and some discussions for further work.

2. Related Work

Information Extraction is viewed as a successful language technology for capturing patient data from unstructured medical texts. The classical rule-based IE paradigm involves extraction of entities after shallow analysis, recognition of references, creation of databases, and filling templates [1]. The integration of machine-learning

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 1–7

approaches, like e.g. the classification of sentences enables recognition of patient features with high precision and recall [2].

Shallow analysis in the IE systems is often based on pattern matching involving cascading applications of regular expressions. Some of these patterns are manually produced and their adaptation to new domain requires much effort. Other patterns are semi-automatically produced by using general meta-rules but they are not too precise [3].

IE is applied in various prototypes which are constructed to perform different extraction tasks from medical documents, including the following ones:

- ***Processing of patient symptoms and diagnosis treatment data***: the system CLEF (Clinical E-Science Framework) extracts data from clinical records of cancer patients [4]; AMBIT acquires Medical and Biomedical Information from Text [5]; MiTAP (MITRE Text and Audio Processing) monitors infectious disease outbreaks and other global events [6]; the system caTIES (Cancer Text Information Extraction System) processes surgical pathology reports [7]; the Medical Language Extraction and Encoding System (MedLEE) was designed for radiology reports and later extended to other domains such as discharge summaries [8]. Other systems are HITEx (Health Information Text Extraction), an open-source NLP system [9] and cTAKES (clinical Text Analysis and Knowledge extraction system) [10];
- ***Building of medical ontologies***: IE is applied for construction of ontology in pneumology in the PertoMed project. The approach is based on terminology extraction from texts according to the differential semantics theory - distributional analysis and recognition of semantic relationships by lexico-syntactic patterns [11]. ODIE (Ontology Development and Information Extraction) is a software toolkit which codes document sets with ontologies or enriches existing ontologies with new concepts from the document set. It contains modules for Named Entity Recognition, co-reference resolution, concept discovery, discourse reasoning and attribute value extraction [12];
- ***Semi-automatic production of clinical guidelines***: [13] presents the systems EviX (Facilitating Evidence-based Decision Support Using Information Extraction and Clinical Guidelines) and LASSIE (modeLing treAtment proceSSes using Information Extraction) that apply IE methods to semi-automatically creation of computer-interpretable clinical guidelines and modeling treatment processes;
- ***Creating databases and digital libraries***: the system EMPathIE applies IE for conjunction of online database from academic journal articles [14]. The system OntoGene extracts semantic relations between specific biological entities (such as Genes and Proteins) from the scientific literature (e.g., PubMed) [15], and PASTA creates a database of protein active sites [16].

Unfortunately the presented IE techniques cannot be directly adapted to our project, because we deal with documents in Bulgarian and many language-processing activities start from scratch. For instance, no Named Entity Recognition has been done for Bulgarian entities in the medical domain; the regular expressions for shallow sentence analysis are constructed in the project for the first time and so on. In this article we present our initial results in automatic processing of PRs in Bulgarian language using manually defined patterns.

3. Patient Status Extraction

The length of PR texts in Bulgarian hospitals is usually 2-3 pages. The document is organised in 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; *(iv)* patient status, including results from physical examination; *(v)* laboratory and other tests findings; *(vi)* medical examiners comments; *(vii)* discussion; *(viii)* treatment; *(ix)* recommendations.

Here we discuss the extraction of patient status data from free text. The relevant PR section contains mostly short declarative sentences in present tense which describe the status of different anatomic organs. At present we do not consider the values concerning Lab and other tests findings. Several organs are referred to in the PRs of patients with diabetes; the text presents the characteristics of 20-30 different anatomic organs and status conditions. The full description might contain more than 45 different status observations. The explanation detailness depends on the status of the corresponding anatomic organs: for some organs only general characteristics are presented, while detailed description is given for other organs which are important for the particular disease. Sometimes organ descriptions are missing and we assume that there is no deviation from the normal status; therefore the system automatically includes certain default values. When an organ description is located in the text, its phrases and sentences are analysed by a cascade of regular expressions.

In order to capture the information we use a terminological bank of medical terms, derived from ICD-10 in Bulgarian language. The International Classification of Diseases (ICD-10) contains 10970 terms. The Bulgarian version of ICD-10 has no clinical extension, i.e. some medical terms need to be extracted from additional resources like a partial taxonomy of body parts, a list of medicines etc. We have compiled a lexicon of medical terminology containing 5288 terms. A lexicon of 30000 Bulgarian lexemes, which is part of a large general-purpose lexical database with 70000 lexemes, completes the necessary dictionary for

morphological analysis of Bulgarian medical text. Another helpful resource is the conceptual model of body parts. It supports the decision how to relate characteristics to anatomic organs when they are presented in separated sentences; this partial medical ontology shows the links between the concepts and points the organs which the attributes refer to. The ontology also supports the dynamic generation of templates if the text contains only partial information about some anatomic organ.

Below we present the typical occurrences of organ descriptions in the PR texts. Let us denote the Anatomic Organs by ***AO*** *(e.g. skin, neck, limbs)*, their characteristics (attributes) by ***Ch*** *(e.g. for skin - colour, hydration, turgor, elasticity etc.)*, the attribute values by ***V*** *(e.g. pale, subicter, decreased, reduced)*, and let ***G*** stands for the general explanation of patient status. Then the status-related text expressions can be grouped into the following categories:

- Description of one ***AO***, all its characteristics and their values presented in one sentence:

AO [-] ['with'/'of'] V1 [Ch1], ['with'/'of'] V2 [Ch2] [and Ch3], ...

"Кожа - бледа, с пепеляв отенък, с намален тургор и еластичност."
(Skin - pale, ash-coloured, with decreased turgor and elasticity)

- Description of one ***AO***, all its characteristics and their values presented in several consecutive sentences:

AO [-] ['with'/'of'] V1 [Ch1]. ['with'/'of'] V2 Ch2 [and Ch3]. ...

"Кожа - бледа. С намален тургор и еластичност. Диабетна рубеоза."

(Skin - pale. With decreased turgor and elasticity. Rubeosis diabetica.)

- Description of one ***AO*** by general characteristics, presented in one sentence:

AO1 [-] [V1,V2] ['with'/'of'] G.

"Кожа бледа с непроменена характеристика. "
(Skin pale with unchanged characteristics.)

- Description of several ***AOs*** having common characteristics and values presented in one sentence:

AO1 and AO2 [-] V1 [Ch1], V2[Ch2], V3[Ch3], ...

"Кожа и видими лигавици - сухи, бледо розови."
(Skin and visible mucoses - dry, light rose coloured.)

- Description of several ***AOs*** having common generally-stated characteristics and values, presented in one sentence:

AO1 and AO2 [-] ['with'/'of'] G.

"Кожа и видими лигавици с нормална характеристика."
(Skin and visible mucous membranes with normal characteristics.)

- Description of several ***AOs*** having different characteristics and values, presented in one sentence:

AO1 Ch11 V11, Ch12 V12, ..., AO2 Ch21 V21, Ch22 V22,..., AO3 Ch31 V31 ...

"Мъж на видима възраст отговаряща на действителната, в добро общо състояние, ало- и аутопсихично ориентиран, КМС-ма правилно развита за възрастта, Р-172 см, Т-63 кг., кожа - розова, с нормална характеристика, добре изразена подкожна мастна тъкан, видими лигавици - розови, влажни. "

(A man on apparent age correspondent to the stated one, in good general condition, allo- and auto-orientation orientation to person, place and time, skeletal muscle system well developed for his age, height - 172 cm, weight - 63 kg, skin - rose, with normal characteristics, well presented hypodermic fat tissue, visible mucoses - rose, moist.)

About 96% of all PRs in our training corpus contain organ descriptions in this format. The more complicated phrases and sentences are analysed by rules for recognising the attributes and their values scope. Some PRs lack descriptions about certain organ attributes. The above-listed six kinds of text formats are recognised by especially prepared rules and regular expressions (taking into account some typical prepositions). The ***AOs*** are usually nouns in basic form and can be identified in the text using the labels of the medical ontology: e.g. "*кожа, щитовидна жлеза, шия, крайници" (skin, thyroid gland, neck, limbs*). The main attributes can be also recognised using the medical terminology lexicon and the medical ontology although some of them have adjacent modifiers like:

- **adverbs** – which express the degree and stage of the symptom, like *умерено, добре, частично (moderately, well, partially)*

- **adjectives**:

(i) some of them express details about certain attributes. For instance, the "skin" characteristic "hydration" can be presented by a variety of wordforms: „*сух*" (*dry*) and others "*възсух, суховат, ..." (slightly dry, rather dry than normal, …)*. In other cases the adjectives express different rates of the attribute values. For instance the noun phrase "*hypodermic fat tissue*" can be modified by adjectives like "*увеличена, леко увеличена, умерено изразена, добре изразена, редуцирана, силно редуцирана" (increased, slightly increased, moderately developed, well developed, reduced, highly reduced).*

(ii) other adjectives represent attributes and participate in the medical terminology lexicon. Sometimes the PRs contain adjectives which are not typical for the medical domain. For instance, the attribute "*colour*" of the organ "*skin*", is usually presented by the adjectives *"бледа, розова, бледо розова, пастьозна, мургава" (pale, rose, light rose, doughty, swarthy)* and non-typical words like *"с пепеляв оттенък, бакърена" (ash-coloured, copper-like).* The successful recognition of these attributes is provided by the large, representative corpus of patient records and the large Bulgarian dictionary with general lexica.

To summarise, in our corpus the patient organs and their features are described in the following way:

- ***General discussion*** – by giving some default value, e.g. *"с непроменена характеристика за възрастта" (with unchanged characteristics typical for the age), "със запазена характеристика" (with preserved characteristics), "с нормална характеристика" (with normal characteristics).* General statements happen relatively often, e.g. the skin status for 26% of the PRs is given by general explanations. For filling in the obligatory IE template fields in these cases, we need predefined default values for the respective organs status. This issue is further discussed below;

- ***Explicit statements*** – the PR text contains specific values. The attribute name might be missing since the reference to the organ is sufficient: e.g. *"pale skin"* instead of *"skin with pale colour"*. The attributes are described by a variety of expressions, e.g. for the *"volume of the thyroid gland"* the value *"нормална" (normal)* can be expressed as *"неувеличена, не се палпира увеличена, не се палпира" (not enlarged, not palpated enlarged, not palpated).* There are several characteristics (about 15% of all attributes) that have numerical values like *"АН в легнало положение 150/110 mmHg, изправено положение 110/90 mmHg." (BP 150/110 mmHg in lying position, 110/90 mmHg in standing position).*

- ***Partial explanations*** – the text contains descriptions about organ parts, not for the whole anatomic organ. For instance, the skin status can be expressed by phrases like *"дифузно зачервяване на лицето" (diffuse redness of the face).* In this case we need some knowledge about the body parts to infer that *face skin* is part of the whole *skin*. Additional fields are added to the obligatory fields of the IE template;

- ***By diagnosis*** – sometimes a diagnosis is given instead of organ description, e.g. "*Затлъстяване от първа степен" (First degree of obesity)* or *"онихомикоза" (onychomicosis).*

It is challenging to extract the patient status description directly from the PRs as it is often not recorded in the patient's clinical notes. Figure 1 presents the percentage of PR texts which discuss five skin characteristics. About 20% of the PRs in our corpus report observations of other skin attributes which are not included in Fig. 1. Much information in the PR text is implicit.

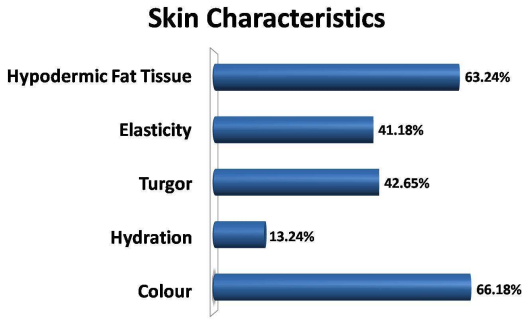

Figure 1. Percentage of PRs with explicit statements about *skin* status

Our approach for negation treatment is based on shallow processing (chunking only) in combination with deep semantic analysis in certain points [17]. The choice of proper templates for recognition and interpretation of the negation influences considerably the system performance. For example we can recognize phrases like:

"без шумова находка" (without thrill finding)
"няма патологични промени" (no pathological changes)
"отсъстват хрипове" (absent rales)
"липсват отоци и варикозни промени" (missing edema or varicose changes)

Figure 2. Obligatory fields in the skin template

In general, the shallow text analysis using regular expressions helps to identify the necessary sentence phrases but the decision for filling in a template is often difficult. Fig. 2/left shows the default template of "*skin*" with its four obligatory characteristics. A general sentence might occur in the text, line "*skin with normal characteristic*", and the IE module has to fill in the default values. Fig. 2/right contains a partially-filled template for a PR containing the sentence "*skin - pale, dry, with decreased turgor and elasticity*". Therefore, we need to invent some methods to 'calculate' the missing status values, using the default

values and reasoning rules. In addition we notice that the values for *turgor* and *elasticity* are not quite independent, so the reasoning rules should be based on certain statistical observations regarding the values' interdependencies.

To study the correlation of values for different organ attributes, the medical experts in the project have developed a scale of *normal*, *worse* and *bad* conditions. Some words from the PRs are chosen as representative for the corresponding status scale and the other text expressions are automatically classified into these typical status grades. Table 1 illustrates the scales for *skin* and gives examples for words signaling the respective status. In fact the regular expressions for shallow analysis map the explicit text descriptions about skin into the chosen categories. In this way all word expressions are turned into numeric values and it becomes possible to study the deviations from the normal condition. The mapping process is not trivial and requires precise elaboration of the regular expressions. Some 95 skin colour characteristics exist in the medical domain, although our present corpus contains less and they all have to be treated by corresponding rules.

Scale	Colour	Hydration	Turgor	Elasticity
0	*rose, swarthy light rose, light swarthy*	*normal*	*good, preserved*	*good, preserved*
-1	*pale, subicterus*	*moderate dehydration, dry*	*reduced*	*reduced*
-2	*icterus, cyanosis, ash-coloured, copper-like*	*severe dehydration*	*poor*	*poor*

Table 1. Status types for skin characteristics

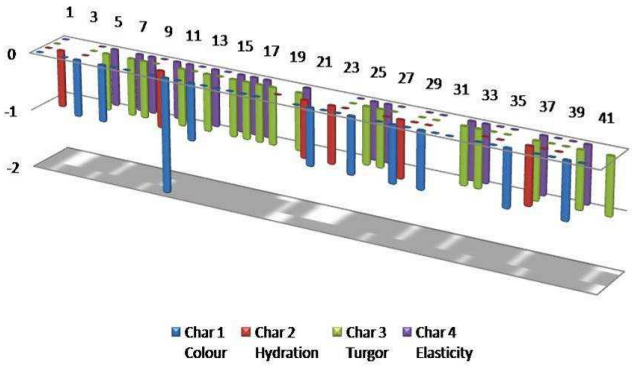

Figure 3. Values correlation in *'skin'* template

Fig. 3 shows the correlations between the values of the fields in *'skin'* template. Each column corresponds to one patient; the values marked there signal the presence of text description in the PR which explains the skin status. We notice that the *turgor* and *elasticity* are usually discussed together.

4. Aggregated Patient Data vs Individual Patient Data

In order to explore the correlations in Fig. 3 we need to analyse the repeated observations on patient status data. Applying standard techniques like Canonical Correlation Analysis and Multiple Regression Analysis is a time and efforts consuming task [18]. Instead of analysing all possible combinations and permutations of values of all characteristics of all anatomic organs, we try to analyse only possible and consistent combinations of values and explore their correlations.

Patient status texts explain not only the current disease; they rather present a complex view which is influenced by all patients' current and previous diseases. That is why it is too difficult to generate the aggregated patient status for a particular disease. First we need to study statistical data about the patient status for each disease. Then we have to select the most typical data and characteristics. For instance if we deal with a disease D_1 we need to explore with high priority the data for patients which have only D_1, but we have to take in consideration also patients with more complex diseases like D_1&D_2, D_1&D_3, D_1&D_2&D_3 etc. There are also sets of diseases that cannot happen together at the same moment, e.g. Diabetes - type 1, Diabetes - type 2, and Diabetes - type 3. Even for patients who have only some disease D_1 the aggregation results are not clear because their status can be influenced from previous diseases which are not mentioned in the present PR texts. In order to avoid inconsistent combinations of diseases we explore only those presented in our corpus.

Below we sketch a data aggregation algorithm which we explore at present in order to complete the picture of patient status data and fill in a special kind of dynamic IE templates. Please note that there could be several PR texts for different visits of the same patient at the hospital.

Algorithm: Let $P=\{p_1, p_2, ..., p_k\}$ be the training set of patients (i.e. text describing patient status data).

Step 1: For each patient $p_j \in P$, $j=1,..,n$ find in the PR text the set q_j of the corresponding diseases of p_j (the mapping is shown at Fig. 4).

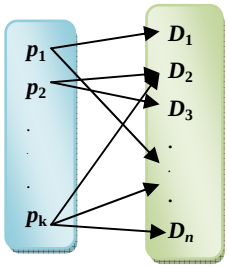

Figure 4. Corresponding diseases for patients in the corpus

Step 2: From the set $Q = \{q_1, q_2, ..., q_k\}$ find $D=\{D_1, ..., D_n\}$ of all diseases for the patients in P;

Step 3: Cluster Q into m classes of equivalence $T=\{T_1,T_2,...,T_m\}$ where the class $T_i=\{q_j \mid q_j \in Q \text{ and } q_j = q_i\}$;

Step 4: Cluster P into m sets $S = \{S_1, S_2, ..., S_m\}$ where $S_i=\{p_j \mid p_j \in P \text{ and } q_j \in T_i\}$;

Step 5: For each set S_i, each anatomic organ and each of its characteristics, calculate the statistical distribution of the attributes among the patients with the class of diagnosis T_i only and compute the most expected (most probable) value of each characteristic. Let us denote it by APD_i (***A***ggregated ***P***atient ***D***ata for the class of diseases T_i).

Step 6: Find $S^{(1)}$ for patients whose diagnosis differ in exactly one disease (one more or less, but have at least one common disease) $S^{(1)}{}_i=\{S_j \mid S_j \in S : |q_j \setminus q_i|+|q_i \setminus q_j|=1\}$ $1 \le i \le n$, $S^{(2)}{}_i=\{S_j \mid S_j \in S : |q_j \setminus q_i|+|q_i \setminus q_j|=2\}$ $1 \le i \le n$ for patients which diagnosis differ in exactly two diseases etc. (See Fig. 5).

Step 7: Refine results for each APD_i by using data from $S^{(1)}$, $S^{(2)}$, ..., $S^{(n)}$. This refinement is done by including patient data from $S^{(1)}$, $S^{(2)}$, ..., $S^{(n)}$. For each set $S^{(m)}$, its APD_i is calculated separately for the patients having T_i. Different decreasing weights $(w_1,w_2,...,w_n)$ are assigned to the APD_i values for the sets in $S^{(1)}$, $S^{(2)}$, ..., $S^{(n)}$.

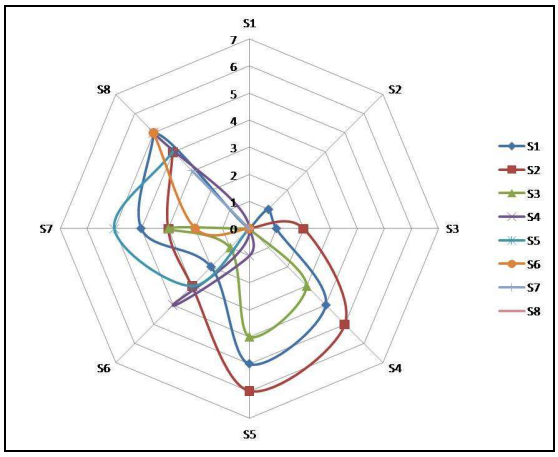

Figure 5. Distances between sets in S

Using the aggregated personal data, we can define templates for each disease and its possible combinations with other diseases. There will be more *expected characteristics* of the anatomic organs – those with probabilities higher than a predefined threshold. Not all combinations of values are consistent and due to this reason we cannot mark directly the default values.

The next step is to produce rules for more possible consistent values for the IE templates. First we rank characteristics according to their number of occurences in the patient status. Then for every possible value of the characteristics, occurring more often, the rest of characteristics values are ranked. The obtained rules will be used for improvement of the generated templates and their obligatory fields with expected values.

This approach for generation of 'intelligent dynamic templates' is under development at present. The objective is to have dynamic templates for each disease, where the expected obligatory fields and their default values can change depending on the information which is filled in for the corpus of PRs in the particular hospital.

5. Evaluation Results and Discussion

We have evaluated the text analysis and the recognition of the basic status scales for the skin attributes (partly explicated in the text as shown in Table 1). In fact our approach has the same objectives like the one presented in [2], where the patient smoking status is classified into 5 categories by selecting sentences which contain the relevant information. As stated above, the negated descriptions are treated as one expression, following a previous study of negative forms in Bulgarian medical patient texts [17].

We have evaluated the extraction progress using a corpus of 197 PRs as a training set and another 43 PR as a test set. The evaluation is done organ by organ since their description in the text are separated and can be analysed independently. There are few PRs without any description of the organ status but they are removed by the evaluation figures.

The first row of Table 2 shows the percentage of correctly recognised attribute descriptions for three anatomic organs: *skin*, *thyroid gland* and *limbs*. The second row of Table 2 shows the percentage of correctly processed PRs.

	skin	thyroid gland	limbs
Correctly recognised characteristics	94.82%	87.35%	84.62%
Correctly processed PRs	94.03%	94.64%	76.75%

Table 2. Percentage of correctly extracted status attributes

The cases of incorrect extraction are due to more complex sentence structures in the PR text which need to be processed by a deeper syntactic analyser. For example:

"Кожа бледа с папулозен обрив по главата, гърдите и гърба, най-вероятно обусловен от калциеви отлагания в меките тъкани. "

(Pale skin with papular rash on the head, chest and back, most likely caused by calcium precipitation in soft tissues.)

Table 2 shows that the simple regular expressions work relatively well and produce enough input for statistical observations and aggregations of patient status data. It is also clear that we need more data to properly develop the algorithms for production of dynamic templates. Currently we have one base template for each anatomic organ; its possible variations depends on the size of the medical ontology branch concerning this anatomic organ. In more complicated cases variations can increase up to 2^{18}.

6. Present Results and Further Work

The article presents on-going work for extraction of patient status data from PR text. In this initial stage we have considered only some relations in the patient's clinical notes. Steps for generation of dynamic templates are sketched. Our present efforts are focused on morpho-syntactic annotation of full sentences in order to train a statistical parser on Bulgarian medical documents. Unfortunately no Named Entity Recognition component is available for Bulgarian, so we have to consider its development too. As further work for negated phrases we plan to refine the chunking algorithm, to enlarge the number of templates and to expand the language and knowledge resources of the system (lexicon, ontology etc.).

In a long run we plan to develop algorithms for discovering more complex relations and other dependences that are not explicitly given in the text, but this is a target for the future project stages.

7. Acknowledgements

This work is a part of the project EVTIMA ("Effective search of conceptual information with applications in medical informatics", 2009-2011) which is funded by the Bulgarian National Science Fund by grant No DO 02-292/December 2008.

8. References

[1] Grishman, R. *Information Extraction: Techniques and Challenges*. In M.T. Pazienza (Ed.), *Information Extraction* (Int. Summer School SCIE-97), Springer Verlag, 1997.

[2] Savova, G., P. Ogren, P. Duffy, J. Buntrock and C. Chute. *Mayo Clinic NLP System for Patient Smoking Status Identification.* Journal of the American Medical Informatics Association, Vol. 15 No. 1 Jan/Feb 2008, pp. 25-28.

[3] Yangarber, R. *Scenario Customization for Information Extraction*. PhD thesis, New York University, New York, January 2001.

[4] Harkema, H., A. Setzer, R. Gaizauskas, M. Hepple, R. Power, and J. Rogers. *Mining and Modelling Temporal Clinical Data*. In Proceedings of the 4th UK e-Science All Hands Meeting, Nottingham, UK, 2005.

[5] Gaizauskas, R., M. Hepple, N. Davis, Y. Guo, H. Harkema, A. Roberts, and I.Roberts. AMBIT: *Acquiring Medical and Biological Information from Text.* In S.J. Cox (ed.) Proc. 2nd UK e-Science All Hands Meeting, Nottingham, UK, 2003.

[6] Damianos, L., J. Ponte, S. Wohlever, F. Reeder, D. Day, G. Wilson, and L. Hirschman. *MiTAP for Bio-Security: A Case Study*. AI Magazine 2002, 23(4), pp. 13-29.

[7] Cancer Text Information Extraction System (caTIES), see https://cabig.nci.nih.gov/tools/caties.

[8] Friedman C. *Towards a comprehensive medical language processing system: methods and issues.* Proc. AMIA Annual Fall Symposium, 1997, pp. 595-599.

[9] Health Information Text Extraction (HITEx), see https://www.i2b2.org/software/projects/hitex/hitex_manual.html.

[10] K. Savova, G. K., K. Kipper-Schuler, J. D. Buntrock, and Ch. G. Chute. *UIMA-based Clinical Information Extraction System*. LREC 2008 Workshop W16: Towards enhanced interoperability for large HLT systems: UIMA for NLP, May 2008.

[11] Baneyx, A., J. Charlet and M.-C. Jaulent. *Building Medical Ontologies Based on Terminology Extraction from Texts: Methodological Propositions*. In S. Miksch, J. Hunter, E. Keravnou (Eds.) Proc. of the 10th Conference on Artificial Intelligence in Medicine in Europe (AIME 2005), Aberdeen, Scotland, Springer 2005, LNAI 3581, p. 231-235.

[12] Ontology Development and Information Extraction tool, https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/Ontology_Development_and_Information_Extraction_(ODIE), version 19 August 2009.

[13] Kaiser, K., C. Akkaya, and S. Miksch. *How Can Information Extraction Ease Formalizing Treatment Processes in Clinical Practice Guidelines? Artificial Intelligence in Medicine*, Volume 39, Issue 2, Pages 97-98.

[14] Humphreys, K., G. Demetriou, and R. Gaizauskas. Two *Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures.* In Proceedings of the Pacific Symposium on Biocomputations, 2000, pp. 505-516.

[15] Rinaldi, F., G. Schneider, K. Kaljurand, M. Hess, C. Andronis, A. Persidis, and O. Konstanti. *Relation Mining over a Corpus of Scientific literature.* In S. Miksch, J. Hunter, E. Keravnou (eds.) Proceedings of the Conference on Artificial Intelligence in Medicine (AIME 2005), Aberdeen, Scotland, 2005, pp. 535-544.

[16] Gaizauskas R., G. Demetriou, P. J. Artymiuk and P. Willett. *Protein structures and IE from biological texts: the PASTA system*. Bioinformatics, 2003 19(1): pp. 135-143.

[17] 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, 2005, Borovets, Bulgaria, pp. 1-8.

[18] Altman, D., *Practical Statistics for Medical Research*, CRC Press, 1991.