Obtaining Status Descriptions via Automatic Analysis of Hospital Patient Records

Svetla Boytcheva
State University of Library Studies and Information Technologies, Sofia, Bulgaria
E-mail: svetla.boytcheva@gmail.com

Ivelina Nikolova, Elena Paskaleva and Galia Angelova
Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria
E-mail: {iva, hellen, galia}@lml.bas.bg

Dimitar Tcharaktchiev
University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev"
Medical University, Sofia, Bulgaria
E-mail: dimitardt@gmail.com

Nadya Dimitrova
National Oncological Hospital, Bulgarian Cancer Registry, Sofia, Bulgaria
E-mail: dimitrova.nadia@gmail.com

Keywords: automatic natural language processing, information extraction, hospital patient record, patient status, template filling, structured representation

Received: November 16, 2009

Abstract. This article describes the automatic processing of medical texts in order to extract important patient characteristics, thus turning the free text description into a structured internal representation. Shallow text analysis is implemented due to the medical language complexity. The paper sketches the information extraction process and discusses the role of domain knowledge in text analysis. The approach to domain model construction is presented. Evaluation results concerning extraction of patient diagnoses and status are summarised.

Povzetek: Predstavljena je metoda za gradnjo semantičnih podatkov o pacientih iz nestrukturiranega besedila.

1 Introduction

Medical patient records are important documents that were created, processed and stored since the ancient times. They keep the patients' diagnoses, treatments, manipulations etc. Nowadays their role is growing together with the increasing potential for collecting, storing and processing of medical information. Much data values are structured by the Hospital Information Systems – for instance, the numeric values of lab tests are automatically entered in predefined fields, and the drugs prescribed to the patient are maintained via the so-called Computerized Physician Order Entry. However, essential findings are traditionally stored as free text descriptions. In this way the automatic text analysis is viewed as an information technology of vital importance, 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, maintenance of adverse drug events, healthcare management and so on. There are major advances in several directions of medical text processing. One important task is to implement tools for automatic extraction and coding of patient-related information with respect to some established classification schemes, such as ICD (the International Classification of Diseases); in this scenario the automatic extraction can provide essential optimisation of health management tasks. Another important objective is to support knowledge discovery in medicine by doing research on disease causes and symptoms, since the automatic text analysis enables searching for effective treatment methods in patient records' texts. In this approach the medical texts are "translated" to internal formalised representations; then inference algorithms can reveal interconnections and regularities between facts and concepts that could remain unnoticed otherwise. Unfortunately most of the medical documents are available as free texts only, which is a major obstacle to the automatic Information Extraction (IE). Despite the difficulties and challenges, however, there is a growing number of industrial systems and research prototypes in

many natural languages, which perform information extraction from patient-related texts. So the application of language technologies to Patient Records (PRs) free text is viewed nowadays as a must in health informatics.

This paper describes an IE prototype which is applied to PR texts in Bulgarian language. The extraction tasks run on anonymised records for hospital treatments of diabetic patients. Section 2 summarises related research dealing with IE from medical texts. Section 3 presents our prototype: the linguistic and conceptual resources and the IE phases for extraction of patient status data. Explicitly-declared domain knowledge enables application of constraining rules and inferences. Section 4 summarises recent evaluation results. The experiments are run within an integrated multifunctional prototype which supports constant collection of new training data. The conclusion and plans for further work are given in Section 5.

2 Related work

Information Extraction is a popular Natural Language Processing (NLP) approach which was proposed in the 1980s as a flexible technology for analysis of domain texts. It extracts only the *relevant* information and ignores the rest, assuming that it is either too difficult to be captured by shallow techniques or consists of irrelevant words (and hence, by default deals with topics which are irrelevant to the problem in question). Relevant information is communicated by relevant words, so there are clear signs where to look and what to analyse in the message. In this way the IE systems are tailored to the extraction of specific facts only, by knowing in advance the words that can signal the entities and relationships of interest. As the overview [1] points out, IE requires "deeper analysis than key word searches but focuses on surface linguistic phenomena that do not require deep inference". In this way IE represents a midpoint between keyword identification and full text understanding. The classical rule-based IE paradigm involves Named Entity Recognition, extraction of entities after morphological analysis, recognition of phrasal expressions and shallow syntactic analysis, recognition of (co-)references, creation of databases, and filling event templates [2].

Recent IE systems typically achieve more than 90% accuracy in Named Entity Recognition, about 80% in template elements construction and about 60% in scenario template production. Most often IE is limited to "the 60% barrier" because of erroneous system choices in the recognition of coreferences between entities and events; another possibility is that this barrier is due to the shallow analysis potential since IE avoids interpretation of implicit relationships and deep inference [1]. In specific domains, however, and with suitably defined IE targets, the automatic extraction features higher precision and recall. Nowadays IE is the common approach to automatic text analysis in biomedicine, but more fundamental research is needed to advance automatic text understanding in principle; there are high expectations that the NLP progress would enable radical improvements in the clinical decision support, biomedical research and the healthcare sphere in general [3].

Current systems for automatic text analysis are usually focused on specific topics only due to domain complexity and the very large number of entities and relationships there. The technology 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]. Other recent systems are HITEx (Health Information Text Extraction), an open-source NLP system [8] and cTAKES (clinical Text Analysis and Knowledge extraction system) [9];
- ***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 [10]. 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, coreference resolution, concept discovery, discourse reasoning and attribute value extraction [11];
- ***Automatic assignment of ICD codes to diagnoses extracted from patient records***: the article [12] summarises the results of the 2007 Computational Medicine Challenge, a competition which was run on anonymised radiology reports. The top coding systems achieved 89% accuracy and the mean was 76,7%. The three top systems processed the negation, hypernyms and synonyms in some way and exploited the UMLS structure [13]. All three systems performed symbolic computations and two of them had in addition some machine-learning components. The overview [12] notes the importance of rule-based text analysis in the coding-oriented NLP tasks.

Current IE systems are often based on shallow analysis by regular expressions and pattern matching. Some patterns are manually produced and their adaptation to new domain requires much efforts. Other patterns are semi-automatically produced using general meta-rules but they are not too precise [14]. The integration of machine-learning approaches, like e.g. classification of sentences, enables recognition of patient attributes with high precision and recall [15].

In addition we should notice the importance of linguistic and conceptual resources and their integration in the IE tasks. The paper [16] discusses the automatic

entity recognition in biomedical texts using a gold standard corpus of 77 English documents with 2124 entities of five types. The authors consider various methods, ranging from dictionary look-up to machine learning approaches, with maximal success of 83% in entities recognition and conclude that dictionary look-up is a promising basic strategy for terminology recognition (which is the technique chosen in our project too). The system MedScan demonstrates the advantages of ontology-driven approaches to medical IE [17]. MedScan processes sentences from MEDLINE abstracts and produces a set of semantic structures representing the meaning of each sentence. In 2003 it extracted information about pathways and molecular networks, so it was tuned to process sentences containing the relevant words in these areas. After parsing, each sentence is represented as semantic frame; an ontological interpreter evaluates the outputs of the NLP component and converts the valid ones into ontological representation. The following accuracy is reported: processed 4,6 million sentences, with 34% correctly parsed sentences but the analysis of errors shows that with larger lexicon and better grammar the system can extract protein function information with precision above 90% [17]. MedScan applies the ontology as a filter to select correct semantic sentence structures and to skip text units which are irrelevant to the target subject.

Most of 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 module has been implemented for Bulgarian entities in the medical domain; the regular expressions for shallow sentence analysis are constructed for the first time and so on. Therefore we need to select some priorities, i.e. which topics are to be treated first. From medical point of view, a significant task is to analyse the hospitalisation effects: what happens to a patient when he or she enters the hospital in status A and leaves it in status B, i.e. how the hospital treatment affects the patient status. Therefore an important activity is the automatic IE of patient status data, especially the diagnoses and the status extraction for organs which are referred to in the PRs of patients with diabetes.

3 Obtaining patient status data from Bulgarian PR texts

In this section we present our approach to extraction of patient status based on cascades of regular expressions. The PR text is split into relevant fragments using a declarative conceptual model of medical entities and relationships among them. We briefly discuss the raw input texts, the linguistic resources, and the domain model construction.

3.1 Corpus of PRs and system resources

The length of PR texts in Bulgarian hospitals is usually 2-3 pages. The record is organised in the following standard 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. So the patient status description is clearly seen in the text, it facilitates the application of IE algorithms.

The PR text contains medical terminology in Latin alphabet (about 1% of all term tokens in our present corpus), sometimes with different transcriptions in Cyrillic alphabet. There are specific term abbreviations both in Bulgarian and Latin (about 3% of the tokens), numerical values (16% of the tokens) etc. In the hospital PRs, complete sentences are rare, since the text contains primarily sentence phrases only. Sometimes there is no agreement between the sentence parts, and the punctuation marks are not properly placed. Further specific problems are due to the highly-inflexional Bulgarian morphology; the terms occur in the text with a variety of wordforms. Our present raw text training corpus consists of 197 anonymised PRs of diabetic patients which contain 166336 word occurrences or 146900 tokens after the elimination of enumerations, tables, indices and repeating wordforms. Actually the training corpus contains some 6400 words, with about 2000 of them being medical terms. The test corpus contains 1000 anonymised PRs of diabetic patients.

In order to capture the patient-related 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 drugs etc. 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. In addition to the lexicons compiled from different sources, the following linguistic and conceptual resources are integrated in the *resource bank* of our system and support the text analysis:

- semi-automatically prepared regular expressions which enable recognition of particular language constructions;
- rules for negation treatment;
- sets of possible and default values for each attribute for each anatomic organ as well as observations about attribute correlations for each anatomic organ. These values – words and phrases - are collected in advance from the representative training corpus of PRs, textbooks, consultations with medical experts etc.;
- templates to be filled in by organ descriptions with associated list of obligatory and optional fields;
- domain model of concepts and relationships relevant to diabetes, including ontology of body parts (see section 3.3) and ontology of diabetes complications which is adopted from the BioPortal resources [18];

- list of drug names and names of relevant medical appliances.

3.2 Shallow analysis of PR texts

The IE procedure for a predefined entity of interest is initiated when a word signalling an entity description occurs in the PR section *(iv)* "patient status". Let us consider an example where the extraction is to be performed for certain anatimic organ (***AO***), where e.g. ***AO***="*крайници*" (limbs) is identified by the morphological analyser. Then the IE system finds in the resource bank the set of ***AO*** characteristics ***Ch***, let in our case ***Ch***={*ankle, leg, peripheral artery, feet, skin, nail*}. Actually the resource bank contains in the domain model the list of all organs related to the chosen ***AO*** which is stored in the domain ontology: especially for lower limbs, the status explanation text can contain details about different limb vein condition, toes, etc. Thus the set ***Ch*** is enlarged to ***Ch'*** for the processed ***AO*** including the other related anatomic organs which potentially can be discussed in the text. Finally the IE system selects from the resource bank the set ***V*** of all relevant characteristics' values. The sets ***V*** and ***Ch'*** contain not only Bulgarian terminology but also Latin terms and their Cyrillic transliteration as well as term abbreviations both in Latin and Bulgarian language. For instance, the reference to the term "*глезен*" (*ankle*) can be found in the PRs also as "*перималеоларни*" (*perimaleolar*) and the term "*периферните артерии*" (*peripheral arteries*) can be represented also as "*a. dorsalis pedis*" and "*аа. дорзалис педис*". Please note that we do not discuss here the possible spelling errors in the text which need to be automatically corrected before the actual IE processing starts; spell-checking should be treated as a technical pre-processing problem in this case.

The further step of the IE algorithm is to determine the scope of the text descriptions where the status of the chosen ***AO*** is presented. Usually this information is given in several consecutive phrasal descriptions or sentences, in a compact manner about one anatomic organ. In our particular example, the system has to decide which adjacent sentences and phrases describe the limb status; the IE analysis for *limbs* will be run only on the selected text fragment. Scoping is made by using the terms and the corresponding concepts in the domain model. There are several rules for scope recognition, let us list two of them:

- ***AO* followed by its characteristics, AO_1 followed by its characteristics, ..., AO_n followed by its characteristics ...**

where AO_n is the first organ in the paragraph not related to the processed ***AO***. For instance:

"*Крайници – отслабени пулсации на a. dorsalis pedis двустранно. Претибиални и перималеоларни отоци. Онихомикоза, tinea pedis. Сукусио реналис – (-) отр. двустранно"* (*Lower limbs – reduced dorsal pedal pulse on both feet. Pretibial and perimaleolar edema. Onychomycosis, tinea pedis. Succusio renalis – (-) bilateral negative*).

The IE system finds in the second sentence a reference to body parts, which are related to limbs - "*претибиални" (leg)* and *"перималеоларни" (ankle).* The third sentence contains the terms *"онихомикоза" (onychomycosis)* - fungal infection of the nails – which is also related to limbs parts and "*tinea pedis",* denoting fungal foot infection. The fourth sentence contains *"succusio renalis"*, which describes a test for pain in the kidney area, and this is a signal that the limbs description is completed.

- ***AO* followed by its characteristics, some characteristics and values not belonging to the set *V* of *AO***. For instance:

"*Крайници - без отоци, варикозни промени, запазени пулсации на периферните артерии, запазени повърхностна, термо и вибрационна чувствителност. Затруднена и болезнена походка, използва помощни средства*" (*Lower limbs – without oedema, varicose changes, palpable peripheral arteries pulse, preserved tactile, thermo and vibratory sensation. Walks with difficulty, algetic gait, uses assistive devices*).

Here the occurrence of the word "*gait*" signals the completion of the limbs description. Further considerations of our heuristic strategy for recognising the irrelevant terms and concepts are given in [19]: if the IE process runs for a term/concept *X*, only concepts linked to *X* by relations *isa*, *part-of*, *has-location* and *associated-with* are considered relevant. In this way the selection of topic-relevant text fragments is done by integral evaluation of linguistic units in the particular input text and corresponding conceptual entities in the domain model.

The shallow syntactic analysis of the selected text fragment is made by application of regular expressions modelling PR phrasal patterns. The IE system finds in the grammatical resources the greediest regular expression that will recognise the maximal part of the sentences selected at the previous step. The system applies the available regular expressions to the text units one by one until a perfect match is found. In case of partial recognition for all of them, the one that fits to the maximal text fragment is selected. We present below two types of regular expressions for *limbs*, out of six types actually used in our IE system. Let us consider the ***AO***s, their characteristics ***Ch*** and their attribute features ***F***. Then the status-related expressions can be grouped into categories, for instance:

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

AO [-] ['with'|'of' F] Ch1, ['with'|'of' F] Ch2, ...

"Крайници без отоци, запазени периферни пулсации, онихомикоза" (*Lower limbs without oedema, preserved peripheral pulse, onychomycosis*).

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

AO [-] ['with'|'of' F] Ch1. ['with'|'of' F] Ch2. ...

"Крайници – без отоци. Запазени пулсации на периферните артерии" (*Lower limbs without oedema. Palpable peripheral arteries pulse*).

About 96% of all PRs in our training corpus contain limbs descriptions in this format, which excludes the application of deeper syntactic analysis at least to the text paragraphs concerning organ descriptions. The above-listed regular expressions are acquired from the training PR corpus, taking into account some typical prepositions and phrasal constructions.

Unrecognised text fragments which contain relevant words are processed by extra rules in order to capture some negative statements. The IE system considers the negated descriptions as one expression, following a study of negative forms in Bulgarian hospital patient records [20]. For instance:

"Крайници - без отоци или варикозни промени, запазени пулсации на периферните артерии" (*Lower limbs – without oedema or varicose changes, palpable peripheral arteries pulse*),

"*Крайници - без отоци, варикозни промени, запазени пулсации на периферните артерии*" (*Lower limbs – without oedema, varicose changes, palpable peripheral arteries pulse*).

In the first sample the negation "without" refers to "*oedema*" and "*varicose changes*" together, but in the second sample the negated word "*without*" refers to "*oedema*" only and staments about the existence of "*varicose changes*" for this patient is positive.

Some more complicated cases are recognised by the rules for resolving the scope of the characteristics and their values. For instance:

"*Крайници - без отоци, липсващи периферни пулсации на аа.дорзалес педис и тибиалес постериор, суха ливидна, атрофична кожа на стъпалата, ливидни студени пръсти, инфектирани разязвявания на дясно стъпало*"

(*Lower limbs – without oedema, absent dorsal pedal and posterior tibial pulses, dry livid atrophic skin of the feet, livid cold toes, infected ulcers of the right foot*).

In this sample we find six different characteristics: the scope of *"absent peripheral pulses"* concerns the "*dorsal pedal arteries"* and *"posterior tibia's artery*". There is only one characteristic for two anatomic organs. Another case is "*dry livid atrophic skin of the feet",* where we have three characteristics for one anatomic organ within one text phase.

Sometimes status descriptions are mising especially when no pathological changes are observed or the examining medical expert relies on tacit knowledge. In a previous paper we have proposed to collect information concening the attribute correlations by making observarions about attribute interdependencies [21]. In this way we can add most probable values in the template fields which have remained empty, because no explicit statements were found in the PR text. To study the correlation of values for different organ characteristics, the medical experts in the project have developed a scale of *normal*, *bad* and *worst* conditions. Some words from the PR texts 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 *limbs* and gives examples for words signalling the respective status. The regular expressions which have been developed for shallow analysis of limbs status map the explicit text descriptions about limbs into the chosen categories. In this way all word expressions are turned into numeric categories, and it becomes possible to study the deviations from the normal condition. The mapping process is not trivial and requires quite precise elaboration and testing of the regular expressions which enable the recognition of the text descriptions. Our approach has similarities to the one presented in [15], where the patient smoking status is classified into 5 categories.

Scale	Ankle	Leg	Peripheral artery pulsation
0	*normal*	*normal*	*normally present*
-1	*(light) swelling*	*oedema*	*reduced*
-2	*solid swelling*	*solid swelling*	*absent*

Table 1: Limbs Characteristics Categorisation

Finally, the IE algorithm has to choose the appropriate template for the captured information, because each template has versions without and with optional fields. Templates are designed after a careful study of the training corpus. For instance, about 99% of the processed PRs discuss explicitly the status of patient *ankle*, *leg* (*ankle* and *leg* status is usually described together) and the *peripheral artery*. Due to the importance of these organs in the status of diabeticians, they are defined as obligatory fields in the limb-status templates. Dynamic generation of template field is possible, to capture the more detailed descriptions of organ status.

Finally, at the last step, the default values are filled in, in case there are obligatory template fields which cannot remain empty. Default values are defined to cope with missing descriptions in the patient's clinical notes. For instance, 77% of the PRs in our training corpus do not discuss explicitly the skin hydration; only 42% discuss the turgor and the elasticity; but 62% discuss the fat tissue and 63% - the skin colour [21]. Therefore, we need to prelist the default status values, to ensure the proper filling of obligatory template fields.

Further details about linguistic particularities of the PR texts, the shallow text analysis and the dynamic template extension are presented in [20], [21] and [22].

3.3 Building domain model to support IE from diabetic patients' PRs

Without making deep inference, the IE applications integrate some kind of ontological resources to consistently interpret the semantic relationships existing among the entities identified in the text. Often these domain models are constructed using standard or widely-used public controlled vocabularies. However, the manual acquisition is time-consuming and non trivial, therefore we need semi-automatic methods for corpus-based term collection and expansion of the controlled vocabulary by conceptual relations. Our domain model has to support the IE tasks as well as further search of

conceptual patterns by providing general and sibling concepts which enable to identify similarities among case histories. In addition we deal with terms in Bulgarian, which are to be mapped to ontological labels in the IE interpretation phase; therefore we need a conceptual resource with labels in Bulgarian. Only the flat nomenclatures ICD-9 and ICD-10 are translated to Bulgarian and can be directly used as a basic terminological lexicon in the IE tasks. Therefore the development of an IE prototype requires conceptual model construction at least for the domain of diabetes.

Starting from the Bulgarian corpus of PR texts and using the Bulgarian terms of ICD-10, we have performed the following automatic steps which facilitated the corpus-based construction of relevant Bulgarian terms:

(i) We have found all corpus wordforms that do not belong to the Bulgarian lexicon of 70000 entries which contains general lexica. Some 75% of them are manually classified as relevant terms (and another 3% are due to spell-errors);

(ii) We have mapped all corpus wordforms to ICD-10 to find domain terms that participate in the nomenclature;

(iii) We have applied a clunker of Bulgarian phrases to the morphologically-analysed PR text which groups single wordforms into phrases. These phrases are mapped to the ICD-10 terms too.

After manual refinement of the joint term collections, we have constructed a list *Diab-Term-Bg* of 1098 terms, which are potential Bulgarian ontological labels in the conceptual model we need to construct. Applying bilingual Bulgarian-English dictionaries and manual correction by medical experts, these terms are translated to English in order to use them as entries for accessing public semantic resources labelled by English vocabulary. Having at hand this list, named *Diab-Term-Eng*, we can search in the UMLS resources, including MeSH, SNOMED, ICD and so on.

The medical nomenclatures, controlled vocabularies and ontologies in UMLS are not readily suited for our purposes. For instance, MeSH (Medical Subject Headings) - the USA National Library of Medicine's controlled vocabulary thesaurus is a polytree, a hierarchical structure containing 22568 descriptors. The top level concepts are labeled by broad categories such as *Anatomy*, *Diseases*, *Organisms*, etc. The MeSH hierarchy is a forest with 16 heads and depth 11. It contains concepts and relations of synonymy, near-synonymy, and closely related concepts. The MeSH thesaurus was initially proposed for indexing, cataloguing, and searching for biomedical documents. Recently MeSH terms are actively used to e.g. improve information retrieval (by query expansion) but it is hard to apply them as NLP ontological backbone, since most concepts have no property-value specifications, and many available properties convey either very general relationships or relationships that are hard to interpret in the NLP context [23]. Therefore we combine automatic extraction of important UMLS fragments and manual reviewing and editing in order to reduce the ambiguity and to assert the conceptual relations needed to support the IE tasks in the diabetes domain.

For mapping English medical terms to UMLS concepts we use the UMLS tool Metamorphosys and the UMLSKS server. In this way we retrieve the term's concepts with their synonyms, definition, semantic types and sub-concepts together with pointers to the different vocabulary sources. There could be several concepts corresponding to a given term, and manual editing is needed to filter the *isa*-hierarchy and tailor it for our application-tailored domain model. We extract and process hierarchies starting from top categories like *Disease or Syndrome* and *Anatomical Structure*. Figure 1 illustrates the adjusted hierarchical structures we obtain after manual editing of UMLS fragments. As background annotation, we store markers pointing to the UMLS resources which are reviewed in the acquisition process.

In addition to the hierarchical refinements, we need to construct the relations among the concepts of interest. UMLS contains two basic relation types: the hierarchical *isa* and the relation *associated_with* with five sub-relations: *physically_related_to*, *spatially_related_to*, *functionally_related_to*, *temporally_related_to* as well as *conceptually_related_to*. The tree of *associated_with* has depth 4 and contains 52 subrelations, some of them shown at Figure 2. For instance, the important *part_of*

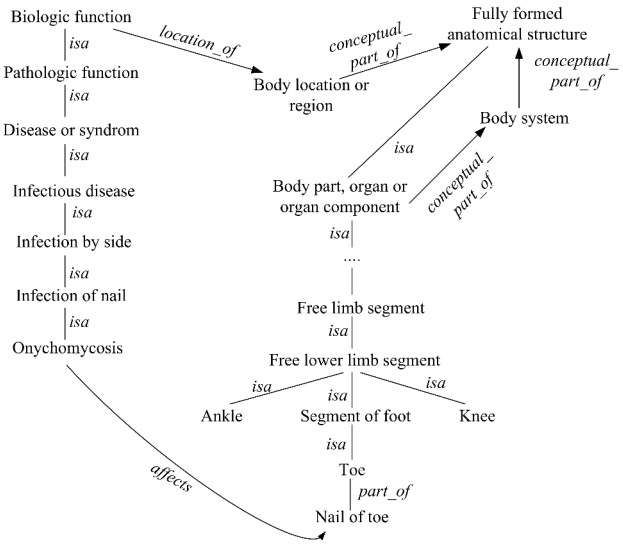

Figure 1. Semantic network for diseases and anatomic organs constructed using automatically extracted UMLS fragments

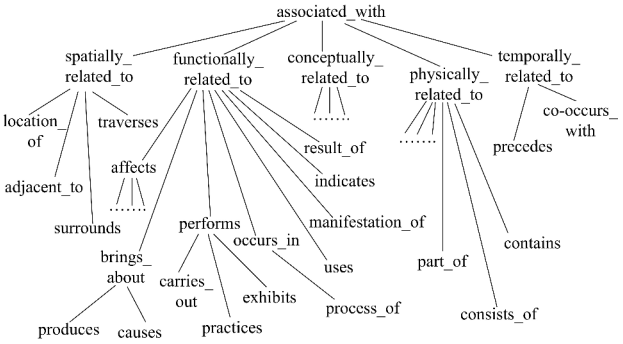

Figure 2. UMLS relations: 28 subrelations of *associated_with*

relation is a subtype of *physically_related_to* and has siblings such as *consists-of*, *contains*, *connected_to* etc. Using the extracted hierarchical structures, we link concepts by UMLS relations. Currently this process is completed for about 300 concepts needed to support the IE of diabetic patient status data. Finally, in our acquisition workbench we provide labelling of domain concepts by the relevant Bulgarian terms. This is necessary, since the IE tasks run on Bulgarian PR texts, and they map input words to domain concept labels during the IE interpretation phase.

Another domain model part concerns the templates where the IE system captures the extracted status data. Usually the IE templates are tables and database entries, but in the medical domain we take into consideration the available archetypes (patterns of standartised structures which normalise the descriptions of various medical artefacts). Archetypes are developed by the openEHR Foundation, an international body which aims at the development of interoperable electronic health records in Europe [24]. They are regarded as an obligatory element of the future EU eHealth framework.

4 Evaluation of the IE prototype

Successes and failures of IE performance are measured by special evaluation exercises which prove the feasibility of the approach to perform partial analysis only, tackling selected entities and relationships. The IE performance is assessed in terms of three classical measures. The *precision* is calculated as the number of correctly extracted entity descriptions, divided by the number of all recognised entity descriptions in the test set. The *recall* is calculated as the number of correctly extracted entity descriptions, divided by the number of all available entity descriptions in the test set (some of them may remain unrecognised by the particular IE module). Thus the precision measures the success and the recall – the recognition ability and "sensitivity" of the algorithms. The F-measure (harmonic mean of precision and recall) is defined as

$F = 2 \times Precision \times Recall / (Precision + Recall)$.

We have developed a prototype which integrates various functions for maintaining the training and test corpus of anonymised PR texts. The IE tasks include browsing and searching functionality, visualisation of the internal templates to the user (see Figure 3) and options for manual editing especially when diagnose codes are assigned [25]. This integrated prototype serves as a convenient unified software environment which is used by project developers and medical experts. We present recent evaluation figures concerning various IE tasks.

At first we summarise the detailed evaluation of patient status IE which is presented in [22]. Table 2 shows the precision, recall and the F-measure of correctly extracted descriptions of the anatomic organs *skin*, *neck*, *thyroid gland* and *limbs*, as well as statements about the patient *age*. We remind that during the analysis and recognition process, the status values are classified as *good*, *fair* and *serious*, which is visually reflected in the interface at Figure 3 by white, yellow and red colours of the respective fields. Green fields contain default values which are automatically filled in for missing obligatory attributes. We see that shallow analysis by regular expressions works relatively well, and the figures shown in Table 2 are comparable to the accuracy of the IE systems presented in section 2. The cases of incorrect analysis are due to more complex syntactic structures in the PR text which need to be analysed by a deeper syntactic parser and semantic processing. Further efforts are also needed to tackle complex language constructions including scope of quantifiers, temporal qualifications etc.

Training set	Skin	Neck	Thyroid gland	Limbs	Age
Precision	95,65	95,65	94,94	93,41	88,89
Recall	73,82	88,00	90,36	85,00	90
F-measure	83,33	91,67	92,59	89,01	89,44

Table 2: Precision, recall and f-measure of extracted patient characteristics

Another important IE task concerns the automatic assignment of disease codes using ICD-10 terms, in order to support the manual coding of patient information and the delivery of health management data. Diagnoses are declared in the PR section *(ii)* "diagnoses of the leading and accompanying diseases". This section contains enumeration of various disease names separated by the punctuation mark full stop. In other words, this section consists of separated, clearly disconnected phrases which are to be mapped to the ICD disease names. In general the diseases in section *(ii)* are not formulated according to the standardised ICD terms, sometimes the disease description might have no common words with the respective ICD term at all. Further mismatches between diseases descriptions in PR texts and the standardised ICD terms are discussed in [26].

The training set for this IE task contains 197 PRs, and the evaluation was performed for a test corpus of 250 unknown PRs. Almost all PRs in the test corpus cite more than one disease per patient, and the number of diseases ranges from 1 to 20. However, when numerous diseases are listed, their phrasal descriptions are often mixed in complex syntax groups; therefore we have performed the evaluation task for 20 test corpus subsets grouping PRs with equal number of diagnoses together. The evaluation results are illustrated by Figure 4 and Figure 5. The *x*-axis of both diagrams represents the twenty PR "test families" consisting of PRs with 1-20 diseases. Figure 4 summarizes the results from the PR perspective. For some PRs, part of the diagnoses are correctly encoded and others are wrong, so Figure 4 shows the ratio of PRs with fully associated diagnoses vs total number of PRs tested. The evaluation can be also made from the perspective of recognised individual diagnoses. Figure 5 presents the ratio of correctly associated codes for diagnoses compared to the total number of diagnoses included in the corresponding PR set.

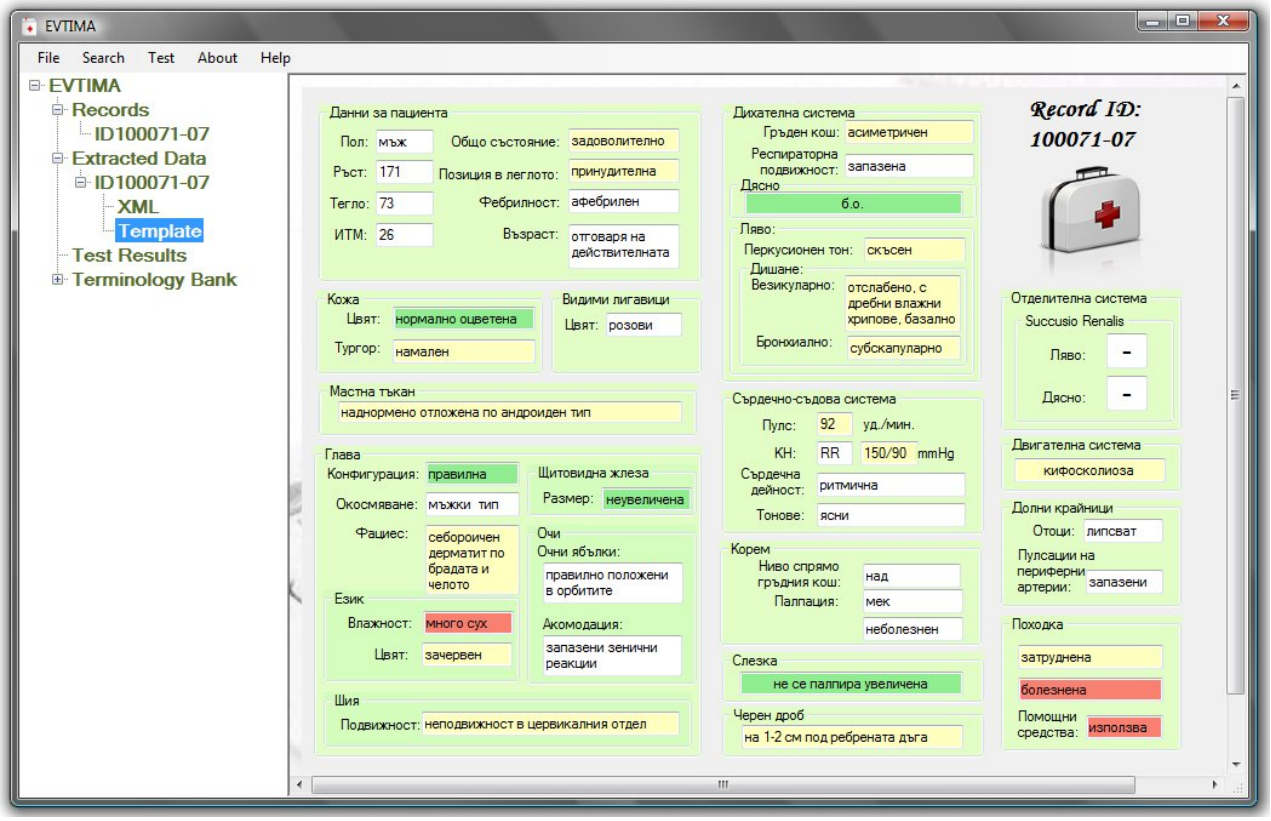

Figure 3: Structured description of patient status data supported by the IE prototype

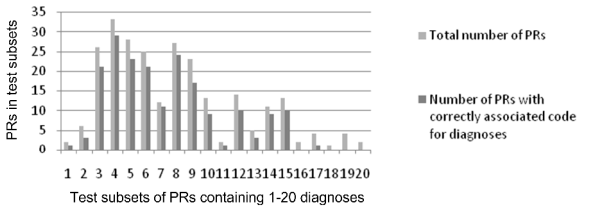

Figure 4. Percentage of PRs with correctly associated ICD-10 codes

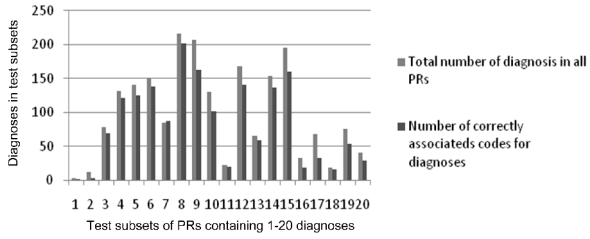

Figure 5. Percentage of diagnoses with correctly associated ICD-10 codes

The evaluation results at Figures 4 and 5 are further explicated at Table 3. Column 2 shows the performance assessment when the whole PR section *(ii)* is submitted to the assigning module as a single text fragment. Column 3 displays the results when the assignment is done phrase by phrase, i.e. every string between two separators in PR section *(ii)* is processed separately. The accuracy is higher when single phrases are considered.

Training set	Sets of diagnoses in PR section *(ii)*	Single diagnose
Precision	81,32	85,73
Recall	76,28	83,96
F-measure	78,72	84,84

Table 3: Precision, recall and f-measure of automatically assigned ICT-10 codes

5 Conclusion

The article describes current results in extraction of patient status data from medical text. It shows the complexity of medical text processing which is due to the complexity of the medical domain and the particularities of the medical texts written in specific, well-established style. The role of explicitly-declared domain knowledge is shown; it supports the information extraction algorithms by providing constraints and inference mechanisms. Construction of domain knowledge resources is a highly expensive, effort-consuming and tedious task, therefore we try to reuse available public resources as much as possible. At the same time the article illustrates the obstacles to build semantic systems in the medical domain: this requires much effort for construction of the conceptual resources as well as the lexicons and grammatical knowledge in case of text processing. Much knowledge in the medical documents is implicit, and its explication in the IE process is a real interpretation challenge.

Despite the difficulties, the paper shows that certain facts can be extracted relatively easily. These promising results support the claim that the Information Extraction approach is helpful for the obtaining of specific medical statements which are described in the PR texts. As future work, we plan to develop algorithms for discovering more complex relations and other dependences among the PR entities.

Acknowledgements

The research work presented in this paper is partly supported by grant DO 02-292/December 2008 "Effective search of conceptual information with applications in medical informatics", funded by the Bulgarian National Science Fund in 2009-2011. The primary PR anonymisation is done by the Hospital Information System of the University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev", part of Medical University - Sofia.

References

[1] Hobbs, J. and E. Riloff (2010) Information Extraction. In: Indurkhya, N. and F. J. Damerau (Eds.), *Handbook of Natural Language Processing*, 2nd Edition, Chapman & Hall/CRC Press, Taylor & Francis Group.

[2] Cunningham, H. (2005) Information Extraction, Automatic. In: Brown K. (Ed.), *Encyclopedia of Language and Linguistics*, 14-Volume Set, Elsevier, Second edition.

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

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

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

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

[7] Cancer Text Information Extraction System (caTIES), see https://cabig.nci.nih.gov/tools/caties, last visited August 2010.

[8] Health Information Text Extraction (HITEx), see https: //www.i2b2.org/software/projects/hitex/hitex_manual.html, last visited August 2010.

[9] Savova, G. K., K. Kipper-Schuler, J. D. Buntrock, and Ch. G. Chute (2008) UIMA-based Clinical Information Extraction System. *Proceedings of LREC-08 Workshop W16: Towards enhanced interoperability for large HLT systems: UIMA for NLP*, ELRA, May 2008.

[10] Baneyx, A., J. Charlet and M.-C. Jaulent (2005) 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)*, Springer LNAI 3581, pp. 231-235.
Ontology Development and Information Extraction tool, last visited August 2010 at https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/Ontology_Development_and_Information_Extraction_(ODIE)

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

[12] *Unified Medical Language System*, US National Library of Medicine, National Institutes of Health, last visited August 2010 at http://www.nlm.nih.gov/research/umls/

[13] Yangarber, R. (2001) Scenario Customization for Information Extraction. PhD thesis, New York Univ., NY.

[14] Savova, G., P. Ogren, P. Duffy, J. Buntrock and C. Chute (2008) Mayo Clinic NLP System for Patient Smoking Status Identification. *Journal of the American Medical Informatics Association*, 15(1), pp. 25-28.

[15] Roberts, A., R. Gaizauskas, M. Hepple and Y. Guo (2008) Combining terminology resources and statistical methods for entity recognition: an evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008), ELRA, May 2008.

[16] Novichkova, S., S. Egorov, and N. Daraselia (2003) MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, Oxford University Press, 19(13), pp. 1699–1706.

[17] BioPortal, http://bioportal.bioontology.org/visualize/13578/Diabetes_Mellitus, last visited April 2010.

[18] Angelova, G. (2010) Use of Domain Knowledge in the Automatic Extraction of Structured Representations from Patient-Related Texts. In: Croitoru, M., S. Ferre, and D. Lucose (Eds.): *Conceptual Structures: from Information to Intelligence*, Springer, LNAI 6208, pp. 14-27.

[19] Boytcheva, S., A. Strupchanska, E. Paskaleva, and D. Tcharaktchiev (2005) Some Aspects of Negation Processing in El. Health Records. In Paskaleva, E. and S. Piperidis (Eds) Proceedings of the International Workshop *Language and Speech Infrastructure for Information Access in the Balkan Countries* held in conjunction with RANLP-05, INCOMA, pp. 1-8.

[20] Boytcheva, S., I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev and N. Dimitrova (2009) Extraction and Exploration of Correlations in Patient Status Data. In: Savova, G., V. Karkaletsis and G. Angelova (Eds). *Biomedical Information Extraction*, Proceedings of the International Workshop held in conjunction with RANLP-09, INCOMA, pp. 1-7.

[21] Boytcheva, S., I. Nikolova, E. Paskaleva, G. Angelova, D. Tcharaktchiev and N. Dimitrova (2010) Structuring of Status Descriptions in Hospital Patient Records. In the *Proceedings 2nd International Workshop on Building and Evaluating Resources for BioMedical Text Mining,* associated to the 7th Int. Conf. on Language Resources and Evaluation (LREC-2010), ELRA, May 2010, pp. 31-36.

[22] Nirenburg, S., M. McShane, M. Zabludowski, S. Beale, C. Pfeifer (2005) Ontological Semantic Text Processing in the Biomedical Domain. *University of Maryland Baltimore County, Institute for Language and Information Technologies, Working Paper 03-05*. Available at http://naboo.ilit.umbc.edu/ILIT_Working_Papers/ILIT_WP_03-05_Biomed_Mesh.pdf, last visited August 2010.

[23] http://www.openehr.org, see *Clinical Models* and *Archetype Authoring,* last visited August 2010.

[24] Boytcheva S., G. Angelova, I. Nikolova, E. Paskaleva, D. Tcharaktchiev and N. Dimitrova (2010) EVTIMA: a System for IE from Hospital Patient Records in Bulgarian. In: Dicheva, D. (Ed.): *AI and Knowledge Societies: Learning, Sharing, Amplifying*, Proceedings of AIMSA-2010, the 14th Int. Conference on Artificial Intelligence – Methodology, Systems, Applications, Springer, LNAI, to appear in September 2010.

[25] Boytcheva S. (2010) Assignment of ICD-10 Codes to Diagnoses in Hospital Patient Records in Bulgarian. In: Alfred, R., G. Angelova and H. Pfeiffer (Eds.). *Proceedings of the International Workshop "Extraction of Structured Information from Texts in the Biomedical Domain" (ESIT-BioMed 2010)*, associated to the 18th Int. Conference on Conceptual Structures (ICCS-2010), Kuching, Sarawak, Malaysia, Published by MIMOS BERHAD, pp. 56-66.