EVTIMA: a system for IE from hospital patient records in Bulgarian

Svetla Boytcheva[1], Galia Angelova[2], Ivelina Nikolova[2], Elena Paskaleva[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
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria, {iva, galia, hellen}@lml.bas.bg
[3] Medical University, Sofia, Bulgaria, dimitardt@gmail.com
[4] National Oncological Hospital, Sofia, Bulgaria, dimitrova.nadia@gmail.com

Abstract. In this article we present a text analysis system designed to extract key information from clinical text in Bulgarian language. Using shallow analysis within an Information Extraction (IE) approach, the system builds structured descriptions of patient status, disease duration, complications and treatments. We discuss some particularities of the medical language of Bulgarian patient records, the role of declarative domain knowledge in the IE process, the architecture and functionality of our current prototype, and evaluation results regarding the IE tasks we tackle at present.

Keywords: biomedical natural language processing, knowledge-based aspects of information extraction from free text, applied AI systems.

1 Introduction

Patient Records (PRs) contain much textual information. Studying hospital PRs in Bulgarian language, which discuss a single hospital episode, one discovers that the most important findings, opinions, summaries and recommendations are stated as free text while clinical data usually support the textual statements or provide clarification of particular facts. Thus the essence of patient-related information is communicated as unstructured text message of one medical expert with addressee another medical expert. In addition 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. However, automatic analysis of biomedical text is a complex task which requires various linguistic and conceptual resources [1]. Despite the difficulties and challenges, however, there are industrial systems and research prototypes in many natural languages, which aim at the information extraction of features from patient-related texts. So the application of language technologies to medical PRs is viewed as an advanced but standard task which is a must in health informatics.

In this paper we present an IE prototype which extracts important facts from hospital PRs of patients diagnosed with different types of diabetes. The system, called

EVTIMA, is under development in a running project for medical text processing in Bulgarian language. It should be classified as an ontology-driven IE system, following the classification in [1]. The article is structured as follows: section 2 briefly presents related approaches; section 3 discusses the prototype and its functionality; section 4 deals with the evaluation and section 5 contains the conclusion.

2 Related Work

When designing our system, its rules for shallow analysis and the training corpus, we have studied carefully the CLEF site [2]. Other systems which process patient symptoms and diagnosis treatment data are: the Medical Language Extraction and Encoding System which was designed for radiology reports and later extended to other domains such as discharge summaries [3]; caTIES [4] which processes surgical pathology reports; the open-source NLP system Health Information Text Extraction HITEx [5], and the Clinical Text Analysis and Knowledge extraction system cTAKES [6]. Interesting and useful ideas about processing of medical terminology and derivation of terminological variants are given in [7]. Negative statements in Bulgarian patient-related texts are studied in [8].

The IE approach arose in the late 80ties as an approach for partial natural language understanding, i.e. extraction of entities and relations of interest without full text understanding (see e.g. [9]). Various IE applications work on free text and produce the so-called templates: fixed-format, unambiguous representations of available information about preselected events. The IE steps are implemented after text preprocessing (tokenisation, lemmatisation) and the task is usually split into several components [10]:

- *Named entity recognition* – finds entities (e.g. nouns, proper names) and classifies them as person names, places, organisations etc.;
- *Coreference resolution* – finds which entities and references (e.g. pronouns) are identical, i.e. refer to the same thing;
- *Template element construction* – finds additional information about template entities – e.g. their attributes;
- *Template relation construction* – finds relations between template entities;
- *Scenario template production* – fills in the event template with specified entities and relationships.

These five steps are convenient for performance evaluation which enables the comparison of IE systems (because one needs intermediate tasks where the performance results can be measured). Recent achievements for English are: more than 95% accuracy in Named entities recognition, about 80% in template elements construction and about 60% in scenario template production [10]. Evaluation results for our present prototype will be presented in Section 4.

3 IE in the EVTIMA system

The system presented here deals with anonymised PRs supported by the Hospital Information System (HIS) of the University Specialised Hospital for Active Treatment of Endocrinology "Acad. I. Penchev" at Medical University, Sofia. The current HIS facilitates PR structuring since the diagnosis, encoded in ICD-10 (the International Classification of Diseases v. 10), is selected via menu. The drugs prescribed to the patient, which treat the illness causing the particular hospital stay, are also supported via the so-called Computerised Provider Order Entry. In this way some important information is already structured and easy to find. However, in the PR discussion of case history, the previous diseases and their treatments are described as unstructured text only. In addition in the hospital archive we find PRs as separated text files, and these PRs consist of free text. Therefore EVTIMA needs to recognise the ICD terms, drug names, patient age and sex, family risk factors and so on. It is curious to note that the list of drugs and medications is supported with Latin names by the Bulgarian Drug Agency, even for drugs produced in Bulgaria [11], but in the PR texts the medications are predominantly referred to in Bulgarian language, so the drug-related vocabulary is compiled on the fly in two languages.

3.1. Medical Language in Hospital PRs

The length of PR texts in Bulgarian hospitals is usually 2-3 pages. The document is organised in the following zones: *(i)* document identification; *(ii)* personal details; *(iii)* diagnoses of the leading and accompanying diseases; *(iv)* anamnesis (personal medical history), including current complains, past diseases, family medical history, allergies, risk factors; *(v)* patient status, including results from physical examination; *(vi)* laboratory and other tests findings; *(vii)* consult with other clinicians and *(viii)* discussion (some of these zones might be omitted in the PRs, see Table 1). Despite the clear PR fragmentation, there are various problems for automatic text processing.

Bulgarian medical texts contain a specific mixture of Cyrillic and Latin terms transcribed with Cyrillic letters, and terminology in Latin. The terms occur in the text with a variety of wordforms which is typical for the highly-inflectional Bulgarian language. The major part of the text consists of short declarative sentences and sentence phrases without agreement, often without proper punctuation marks. There are various kinds of typos in the original text which cannot be properly tackled within our research project. In the evaluation we consider only normalised texts with standard abbreviations and without spelling errors, because we aim at a research study. Another PR text particularity is that specific descriptions are often omitted since the medical experts consider them insignificant or implicitly communicated by some other description. As reported in [12], only 86% of the PRs in our corpus discuss explicitly the patient status regarding skin colour, 63% - fat tissue, about 42% - skin turgor and elasticity, and 13% - skin hydration. So default values have to be assigned to many IE template slots.

Zones/doc	1	2	3	4	5	6	7	8
No. docs	1	0	0	3	11	79	422	484

Table 1. Number of zones in 1000 PRs

3.2. Shallow Rules for Analysis

The specific nature of the PR language does not allow for deep syntactic analysis; therefore we use only shallow rules for feature extraction at present. Acquisition of extraction rules for features like age, sex, illness and illness duration, was performed semi-automatically. Phrases containing the focal terms were selected with window up to 5 words, left and right. These phrases were clustered and common rules were created. However the features describing organ conditions have more complex representation in the text hence more complicated rules for their extraction are needed. These rules first fix the area where the organ description starts and are applied until some feature signals the end of the organ description. We present examples of rules (regular expressions) matching the characteristics of the skin/limbs. AO stands for anatomic organ, F for features, Ch for feature characteristics, G for generally accepted characteristics value (eg. Default value like ”with preserved characteristics”).

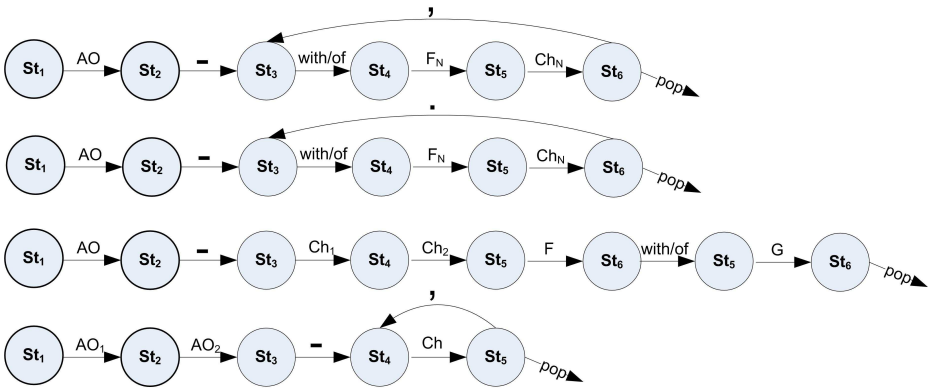

Figure 1. Rules for feature extraction.

About 96% of the PRs in our corpus present organ descriptions as a sequence of declarative nominal phrases, organised into short sentences. In this way the shallow analysis of phrases and sentences by cascades of regular expressions, which are illustrated at Fig.1, proves to be successful approach for structuring the statements concerning patient status. For each particular body part of interest, there is a predefined template, where text units are stored as conceptual entities during the domain interpretation phase. There is some default sequence of discussing the patient status which is kept in all PRs; this genre particularity also helps to design the IE

templates where the text description of patient status is to be structured. Evaluation results are presented in section 4.

In this paper we discuss a project-specific usage of domain knowledge: when a certain particular template is selected for filling and the IE system looks for template elements, the ontology provides constraints and helps to determine a text fragment. Let us consider some examples of PR fragments discussing the patient status:

Sample text 1: Глава – без патологични изменения. Очни ябълки – правилно положени в орбитите, без очедвигателни нарушения. Език - суховат, зачервен. Шия – запазена подвижност. Щитовидна жлеза не се палпира увеличена, пресен несекретиращ оперативен цикатрикс. Нормостеничен гръден кош, ...

Head – without pathological changes. Eyes correctly placed in the eye-sockets, without disturbances in the eye-movements. Tongue – dry, red. Neck – preserved mobility. Thyroid gland does not palpate enlarged, fresh non-secreting operative cicatrix. Normostenic thorax, …

Here the IE system finds the term *"head"* in the first sentence and runs the IE process in order to extract *head's* status, which is to be structured in the slots of a predefined template. The second and third sentences contain terms referring to body parts, linked to *"head"* by the *"has-location"* and *"part-of"* relations - *"eye"* and *"tongue"* which is located in the *head's* part *"mouth"*. Mapping these terms to the concepts and relations in the ontology, the IE system considers sentences 1-3 as status descriptions of *"head"*, *"eye"*, and *"tongue"* correspondingly. The fourth and fifth sentences contain the terms *"neck"* and *"thyroid gland"* which usually appear in consecutive sentences. The *"neck"* is not *head*'s part and therefore, the IE system considers the fourth sentence as a beginning of new body part description. The *cicatrix* in the fifth sentence refers to *"neck"* despite the fact that it is mentioned together with the *"thyroid gland"* in the same sentence. The discussion continues by presenting the *thorax* status in the sixth sentence which signals focus shift to another body part. Usually new descriptions start in another sentence but all statements are mixed into one paragraph. Other examples are:

Sample text 2: Eyes correctly placed in the eye-sockets, slow pupil reactions. Strong vision impairment. Neck – preserved mobility. Thyroid gland enlarged IA stage. Lymph nodes do not palpate. Emphysematous thorax, …

Sample text 3: Head – without pathological changes. Eyes correctly placed in the eye-sockets, normal pupil reactions. Bilateral exophthalmos, without disturbances in the eye-movements. Neck – preserved mobility, palpable veins. Thyroid gland not enlarged. Vesicular breathing, …

In the sample text 2, the "*lymph nodes*" mentioned in the fifth sentence will be interpreted as "*neck lymph nodes*" because this sentence is located immediately after the discussion of "*neck*" and "*thyroid gland*" but "*lymph nodes*" are not part of the "*thyroid gland*". The occurrence of terms like "*thorax*" and "*breathing*" marks the beginning of new body part discussions.

3.3. Present Functionality

The EVTIMA system is designed as a stand-alone desktop application to be used by clinicians and knowledge engineers who want to study the textual data in the patient records. It is created as a scientific prototype and does not have ambitions for an industrial application. Please note that due to medical privacy and confidentiality conditions, we refrain from making a publicly available web-demo of EVTIMA.

The present system is able to work on single and multiple PRs. The implemented functionalities are the following:

- exploring and updating of the terminology bank;
- automatic segmentation of file/s in zones (as described in section 3.1);
- feature and relation extraction from the separate zones by text analysis;
- side-by-side document comparison;
- retrieval of analised documents by given features;
- export of analysises.

EVTIMA is composed of several interrelated modules. It has a *terminology bank* where all vocabularies are kept and used for further IE tasks.

Another module is dealing with the text processing of the PRs. Once a file is loaded, the first task to be performed is automatic segmentation of the input in its semantic zones as described in section 3.1. The zones are to be found by keywords represented by regular expressions. The following text analysis operations are done separately on the so formed parts of the documents. Thus different rules are built for working on distinct areas of the text and for capturing various classes of features.

So far recognition modules are available for the diagnosis, anamnesis and status zones. The user may run automatic recognition of the following features *age*, *sex*, *diabetes type*, *diabetes term*, *diagnosis*, *skin condition*, *limbs condition*. Sample recognition rules are shown at Figure 1.

The recognition can be done automatically and manually. When the automatic mode is chosen EVTIMA picks the most probable value for each feature. In manual mode the system offers for each predefined feature possible characteristics and their values, it is up to the user to chose which is the correct one to be stored in the analysis. Thus the system supports semi-automatic annotation of the text which is also used to validate the performance of the fully-automatic one. All saved analyses are available in XML format and could be later exported. Exports can be done in two different ways - they are either original texts enriched with annotation information or extracts of structured information captured during the analysis. An example of the second one is a template for patient status shown in Figure 2.

For each PR the system instantiates a new template object whose fields are to be filled after analysing the text. Each feature is presented as the relation of its characteristics and their values. In the future, the conjunction of the recognised features should represent the case summarisation and generalisation. The feature relations' arity could be 1, 2 or more and they can have nested relations. An example for this internal representation is shown at Figure 2. The feature characteristics preserved in the template are normalised.

A retrieval engine is available for document search within the repository of analysed PRs. Search criteria are the features and their characteristic values.

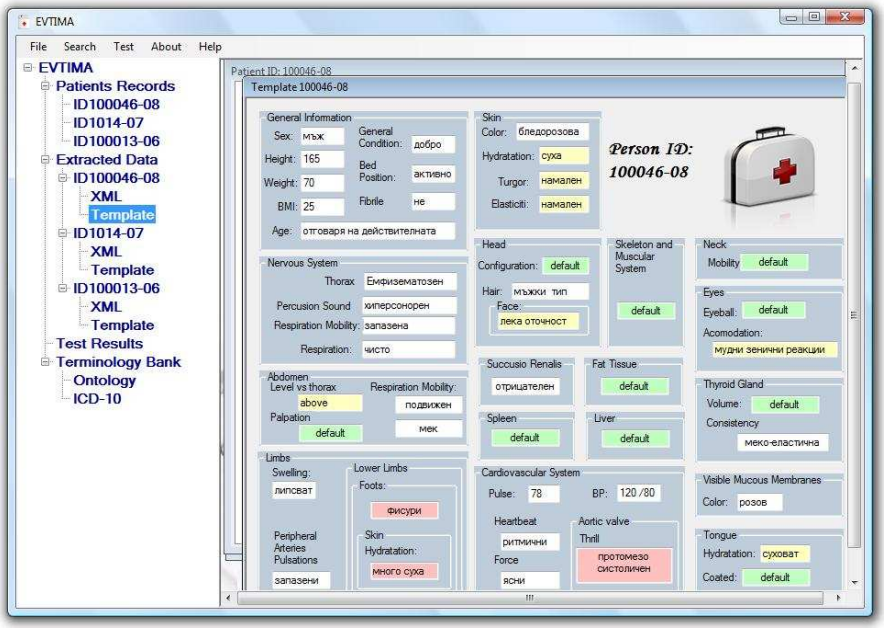

Figure 2. An IE template filled with extracted features of the patient's status.

4 Evaluation

Usually the Information Extraction performance is assessed in terms of three measures. The precision is calculated as the number of correctly extracted features, divided by the number of all recognised features in the test set. The recall is calculated as the number of correctly extracted features descriptions, divided by the number of all available features 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

$$F = 2 * Precision * Recall / (Precision + Recall)$$

The segmentation and extraction of attributes were evaluated using a corpus of 197 PRs as a training set and another set of 1000 PRs as a test set. From the test set we have randomly extracted 200 PRs on which we run the evaluation.

4.1. Segmentation task

The segmentation task even though looking as a more formal one, has been the first obstacle to overcome before analysing the PRs. Markers signaling the beginning of each section are rarely overlapping with other expressions in the PRs which allows for

a high precision of the task. The PRs have been written by differnt clinicians without using a standard form and due to this fact often some of the sections are merged and this results in lower recall. The detailed precision and recall figures are shown in Table 2. The correct identification of these sections is of major importance for the IE tasks to be performed after the segmentation.

Both precision and recall are quite high for this task due to the comparatively strict zone mark-up in the PRs. It was interesting to notice that except from the standard keyterms, zone mark-up is strongly tight to the authors' writing style.

Zone	*Precision* (%)	*Recall* (%)
#1	97.47	88.08
#2	97.43	99.48
#3	97.35	96.42
#4	98.49	99.50
#5	98.49	99.50
#6	98.66	98.28
#7	90.02	87.29
#8	90.19	99.80

Table 2. Zone segmentation precision and recall

4.2. Information Extraction task

In a previous paper we have evaluated the recognition of skin characteristics [12]. These results are presented here, together with additional evaluation data, to allow for a more complete assessment of our IE task. Table 3 summarises the results.

Since we are using nested rules for capturing features, spelling errors in (or lack of) expressions which are to be matched on an upper level may prevent the further matching of the rules in the nested fields. We have noticed that this was one of the reasons for the comparatively low recall value of the diabetes duration, because it is dependent on the match of the diabetes acronym and diabetes type recognition. The recall for the feature sex is surprisingly low. This can be explained with the fact that there were too few samples in the anonymised training corpus and when testing on a new dataset, the available rules could not capture the relevant features. The main reason are the new author styles encountered in the enlarged corpus. These inconsistences are covered by continuous update and adjustment of the IE rules for shallow analysis.

The system design and current functionalities have been also tested by clinicians and they have given their positive opinions regarding the direction of development. Medical experts are attracted by the functionality to compare two templates side-by-side and explore dependencies which are not easy to find by reading the text of single documents. Actually, working with the EVTIMA prototype stimulates the medical professionals to invent further functionality which can be useful in the context of their HIS. So EVTIMA is under development as a joint effort of computer scientists and medical users.

Feature	*Precision* (%)	*Recall* (%)	*F-measure* (%)
Age	88.89	90.00	89.44
Sex	80.00	50.00	61.54
Diagnose	98.28	96.67	97.47
Diabetes duration	96.00	83.33	89.22
Skin	95.65	73.82	81.33
Neck	95.65	88.00	91.67
Thiroid glant	94.94	90.36	92.59
Limbs	93.41	85.00	89.01

Table 3. Evaluation of IE accuracy for patient characteristics

5 Conclusion

Our system is the first one which supports medical text mining for Bulgarian. So far we have evaluated its performance on several IE tasks dealing with the anamnesis and patient' status zones of the PRs. The results are promising and show that partial analysis is a successfull approach and needs to be developed further. EVTIMA serves as well as a tool for semi-automatic annotation of the recognised features. It supports different output formats which facilitate the further processing of the obtained analysed structures. We envisage an extend of its IE capabilities to all PR zones and temporal information exctraction which will help building chronicles. Further exploration of the negative expressions and their semantic interpretation in the internal representation is also one in our future tasks. Our system will be successful if it can serve clinicians and knowledge engineers, by offering explicit and inferred knowledge and dependencies which are not directly obtainable from a single document in free text format.

There are many aspect of development and elaboration of our IE approach which have to be explored in the future. Medical patient records contain complex temporal structures which are connected to patient's case history and the family history. It would be useful for a hospital information system to monitor different hospital episodes of the patient thus supporting temporal information as well [14]. Another interesting question is related to the automatic recognition of illness duration and the periods of drug admission. We plan to develop algorithms for discovering more complex relations and other dependences, which is a target for our future work.

Acknowledgements. The research work presented in this paper is 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.

References

[1] Spasic, I., S. Ananiadou, J. McNaught, and A. Kumar. 2005. *Text mining and ontologies in biomedicine: Making sense of raw text*. Oxford University Press, Briefings in Bioinformatics 2005, Vol. 6(3), pp. 239-251.

[2] Clef Clinical E-Science Framework, 2008. Univ. of Sheffield, http://nlp.shef.ac.uk/clef/.

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

[4] caTIES Cancer Text Information Extraction System, https://cabig.nci.nih.gov/tools/caties, 2006.

[5] HITEx Health Information Text Extraction. 2006. See https://www.i2b2.org/software/projects/hitex/hitexmanual.html.

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

[7] Valderrábanos, A., A. Belskis, and L. I. Moreno. 2002. *Multilingual Terminology Extraction and Validation*. In Proc. LREC 2002 (3rd Int. Conf. On Language Resources and Evaluation), Gran Canaria 2002.

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

[9] Grishman, R. and B. Sundheim. *Message understanding conference - 6: A brief history*. In: Proceedings of the 16th International Conference on Computational Linguistics COLING-96, Copenhagen, July1996.

[10] Cunningham, H. *Information Extraction, Automatic*. Elsevier, Encyclopedia of Language and Linguistics, 2005, available at http://gate.ac.uk/sale/ell2/ie/main.pdf, last visited April 2010.

[11] BDA Bulgarian Drug Agency 2010. see the site http://www.bda.bg/index.php?lang=en.

[12] 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. Borovets, Bulgaria, September 2009, pp. 1-7.

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

[14] Boytcheva, S. and G. Angelova. *Towards Extraction of Conceptual Structures from Electronic Health Records*. In: Rudolph, S., F. Dau, and S. O. Kuznetsov (Eds.): Proceedings of the 17th Int. Conf. on Conceptual Structures (ICCS'09), July 2009, Moscow, Russian Federation. Springer, Lecture Notes in Artificial Intelligence 5662, pp. 100–113.