Towards Extraction of Conceptual Structures from Electronic Health Records

Svetla Boytcheva[1] and Galia Angelova[2]

[1] Department of Information Technologies, State University of Library Studies and Information Technologies
119, Tzarigradsko Shose Blvd., 1784 Sofia, Bulgaria,
svetla.boytcheva@gmail.com
[2] Institute for Parallel Processing, Bulgarian Academy of Sciences
25A Acad. G. Bonchev Str., 1113 Sofia, Bulgaria,
galia@lml.bas.bg

Abstract. This paper presents the general framework and the current results of a project that aims to develop a system for knowledge extraction and knowledge discovery in the texts of Electronic Health Records in Bulgarian. The proposed hybrid approach integrates language technologies and conceptual modeling. The system generates conceptual graphs encoding the patient status history, which contain cause-effect templates for each of the patient's diseases and symptoms as well as links to graphs for records of patient's relatives. We describe simple inference in the generated graphs resource bank. Some experiments and their evaluation are presented in the article.

1 Introduction

The first known medical record was developed by Hippocrates in the fifth century B.C. He prescribed two goals when documenting the patient status in natural language: *(i)* A medical record should accurately reflect the course of disease and *(ii)* A medical record should indicate the probable cause of disease. These goals are still appropriate today and most of the patient documentation is still kept in natural language as free unstructured text. However, Electronic Health Records (EHR) systems provide additional functionality, such as interactive alerts to clinicians, interactive flow sheets and tailored order sets, automatic calculation of the price of the medical treatment etc., all of which cannot be supported in paper-based archives [8].

In many countries most of the medical information is available only in textual form. This makes its automatic processing a very difficult task. So to say, this information is "locked up" in paper documents, files or databases in formats which are not suitable for automated processing [4]. Great efforts have been made to translate this information into certain (semi-)structured representations; the activities always include manual or automatic information extraction from free texts. The main difficulties to structure medical information are the complexity of the domain as a whole, the complex medical language, and the variety of

practices for including text descriptions in EHR, which are too specific for different countries and different languages. Many language processing systems, which extract and codify information from EHR in English have been developed. They reflect at least two principally-different views to patient-related texts. *The first approach* is automatic extraction of information concerning patients diagnosis, treatment, manipulations etc. and automatic coding of this information with respect to some established classification schemes, which are issued by financing or statistical institutions. There are large terminologically-based nomenclatures such as SNOMED (the Systematized Nomenclature of Medicine) and ICD (the International Classification of Diseases). These large terminology collections are unified classification systems, translated in many languages, and support the health management and health statistics. Recent critics to SNOMED explicate some conceptual shortcommings which prevent its application as medical ontology in semantic systems [10]. It remains unclear whether the same kind of terminology-based ontologies can support all principally different applications, which are build on top of medical information extraction. Regarding the extraction precision, the leades in the fileds report successful recognition of the complex medical terminology up to 80% even for English [11]. *The second kind* of prototypes is oriented to medical research and knowledge discovery in medicine. It reflects the AI view to text understanding in order to make inferences, to discover interconnections between facts and concepts which could remain unnoticed otherwise, to spot previously unknown regularities and to develop new medical theories. Most prototypes of this kind are developed for English because they exploit large archives of medical abstracts in English. Practically there are no developments for lesser spoken and minor languages.

Here we present the first steps towards buidling a system for automatic extraction of medical texts in Bulgarian. This research effort is made in a project, supported by the Bulgarian National Science Fund in 2009-2011, which is based on previous achievements in processing Bulgarian morphology using large linguistic resources. In this paper we discuss briefly the general ideas behind the project and present the results of its first steps - design and implementation of the Relations Analysis Module and the Conceptual Graphs generator. Some previously developed components for processing the negation in medical patient records in Bulgarian [1] are extended and integrated in the current system.

This paper is structured as follows. Section 2 overviews briefly some related research and discusses basic language technologies which are used for Information Extraction (IE) in the medical domain. Section 3 describes the general project ideas and introduces a sketchy view to the system architecture. Section 4 presents the Relation Analysis Module and the main types of relations which are automatically recognized at present. Section 5 presents the module for generation of logical forms of Conceptual Graphs (CG) using the templates that are filled in by EHR data. Examples and assessment figures describe the current experiments. Section 6 contains some discussions and the conclusion.

2 Related work

A good overview of current EHR systems and their functionality can be found in [8]. We focus only on the natural language texts in EHR assuming that they are available in certain integrated hospital information system. Another good overview, comparison and evaluation of language technologies which extract information from EHR is given in [3, 2]. The white paper [11] presents recent industrial developments in the field.

Several language technologies are used to extract and codify EHR information. As we said above, the most successful applications run for English due to many reasons, among them the simple morphology. The main approaches include:

- Deep parsing of whole sentence in order to constuct detailed parsing trees and to process further the sentence semantics;
- Partail analysis of sentence segments or local phrases in order to fill in templates and to search for some specific relations and keywords. The main subtasks in this process are:
  - using a shallow parser that captures relations between noun phrases (NPs) [6]. The parser extracts relations between all NPs regardless of their type. Then it searches patterns in the text which are based on English closed-class words - i.e prepositions (*by, of, in*), negation, conjunctions (*and, or*) and auxiliary or modal verbs. The extracted relations can contain up to five arguments: relation negation, left-hand side, connector modifier, connector and right-hand side.
  - searching for cause-effect realtions. This approach was used in [5] to identify and extract cause-effect information that is explicitly expressed in medical abstracts in the Medline database. The system is based on tree-like patterns that indicate the presence of a causal relation in sentences, and which parts of the sentence represent correspondingly the cause and the effect. The patterns are matched to the syntactic parse trees of the sentences. Thus parts of the parse tree are extracted as NPs reffering to the cause or the effect.
  - searching for treatment relations [7] using both semi-automatic and manually constructed linguistic patterns which enable the discovery of treatment relations. Mining for 'association rules' is applied to sample sentences containing both a disease concept and a reference to drugs, to identify frequently occurring word patterns and evaluate whether these patterns could be used to identify treatment relations in sentences.
- Combining several language technologies in a pipe-line environment - e.g. in MedLEE (A Medical Language Extraction and Encoding System [9]).

Specific natural language processing tools are developed to ensure the proper anonymisation of patient records [12] by removal of named entities and replacing them by pseudonyms. Some prototypes deal with the essential problems of negation in medical patient records [13], among them there is a module for processing negation in Bulgarian medical texts [1].

3 Project Settings

The suggested system will extract from text the information needed for automatic generation of a Patient's Chronicle - symptoms and diagnosis. Based on the ideas of granularity shift using CG type definitions, type contraction and type expansion [14] and applying inference rules, some more general statements regarding the patient status will be produced. These 'general' graphs will not deal with the single words and concepts in the EHR personal records but will allow for summarisations of the patient information in more general terms which are used by medical professionals when they describe medical knowledge. The whole conceptual archive will support knowledge discovery in medicine. Today we see it as a hypercube of conceptual graphs, corresponding to patients' EHR and their generalisations. There will be connections between the nodes of different patient graphs in case of family relations. This very challenging and ambitious task includes much research to be performed in several years. At the present moment we can discuss only the Information Extraction (IE) solutions and the generation of conceptual graphs which capture the factology of the individual patient records.

As usual in natural language processing of raw documents, the input medical resources are really problematic - texts with specific abbreviations, numerical values of analyses and clinical test data, medical terminology in Bulgarian in Cyrillics and in Latin (using both the Cyrillic and the Latin alphabets), spell-errors with one or two wrong symbols per word, specific language style of the medical professionals and so on. All these obstacles together are not easy to overcome. Another essential problem is the rich temporal structure of the patient descriptions which prevents the application of standard language processing techniques. Fortunately, we rely on stable modules for morphological analysis, very large morphological dictionaries of Bulgarian and well-studied technologies for corrections of spelling errors, which encourages us to approach the automatic processing of raw medical texts as they are stored in a hospital information system. The test set of available epicrises contains some 8000 words and most of them are included in the very large lexicons of general Bulgarian vocabulary which supports the morphological analysis. So in principle the project is equipped with background resources and tools for natural language processing. Previous research on the negation in Bulgarian medical texts reveals some typical language constructions, specific features of the negation scope and suggests solutions for their processing [1]. Needless to say, the expactation is that the automatic IE will work with partial success and many details (expressed indirectly or by wrong words) will be missed in the texts. But IE success of more than 70% will enable the development of an useful conceptual archive which will provide a good basis for knowledge discovery and conceptual search.

The design of the IE system in the project is strongly influenced by EHR specific structure. The textual part of the EHR in Bulgarian has average length of 2-3 pages and 11 predefined and ordered sections: Personal data, Anamnesis, Status, Examinations, Consultations, Debate, Treatment, Treatment results, Recommendations, Working abilities, and Diagnosis.

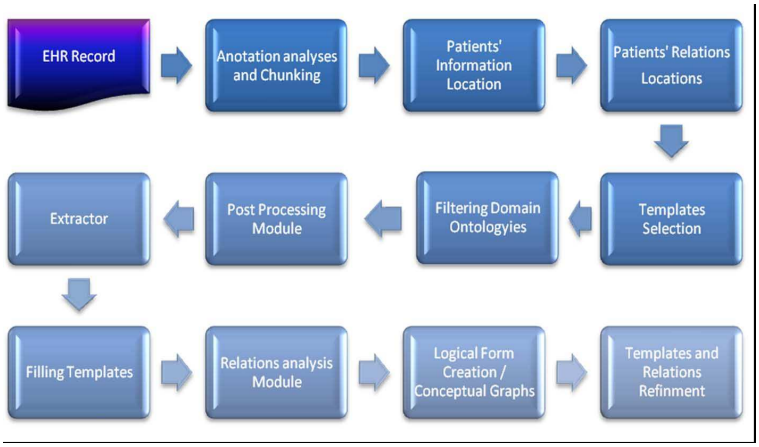

Fig. 1. Pipe-line architecture of modules for information extraction from medical texts

The architecture of the IE component is shown at Fig.1. It contains the following modules:

- AC: Annotation analysis and chunking
- Patients' Data Module
- Patients' Relations Module
- Templates Selection Module
- Post Processing Module
- Extractor
- Filling Templates Module
- Relation Analysis Module
- Logical Form Generator / Conceptual Graphs Generator
- Template and Relations Refinement.

At the first step, each EHR is splits to its 11 sub-topics by the Annotation and Chunking module. The annotation process is based on morphological analysis using a lexicon of 30 000 lexems. The common Bulgarian vocabulary is expanded by medical terminology and specific words which are met in the test EHR corpus. For each EHR word, the module finds its basic form (lexem) with the associated lexical and grammatical features. Chunks are sequences of words that form syntactic groups or sentence phrases. They are recognized by rules, which take into account the morphological features of the words and their mutual position. The module recognizes mostly nominal chunks (NPs) and outputs tagged text and some NPs.

The Patient data are extracted from the Personal data section (taking into account the pseudonymisation) and the system searches the conceptual archive for previous records about the same patient. If any is found, the system creates a new node in the patient graph and includes a pointer to the previous EHR

data according to the patient's chronicle. If no previous data for this patient are found, the system generates a new graph. The Patients' Chronicle graph contains nodes, which are slots of templates full of patient information collected in different time periods (Fig. 2).

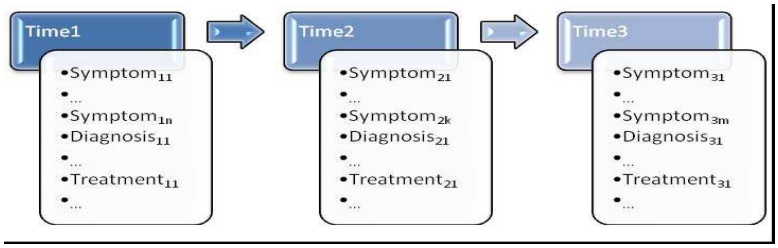

Fig. 2. Patients' Chronicle Relations - Each node is a graph of slots-templates

Patients' Relations Module is responsible for identifying the relatives of the patient. If no information for some of them is available, but the EHR states explicitly that the current patient status results from the illness of some particular relatives, the system creates empty nodes for those "virtual" patients and fills in the corresponding information there. Pointers are included to connect the node of the current patient to the nodes of his/her relatives.

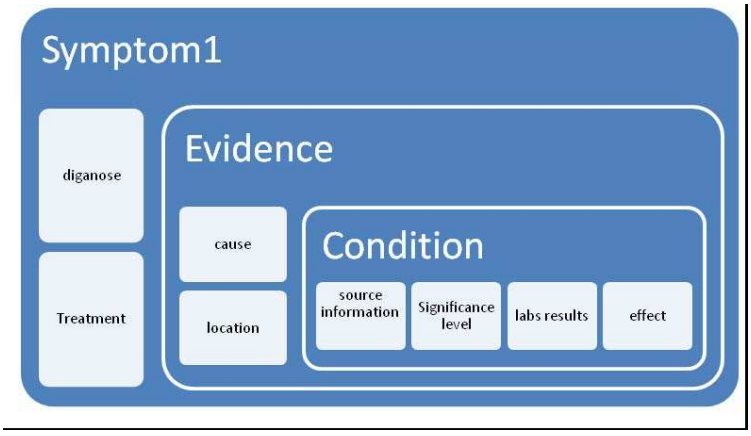

Fig. 3. Patient Status Template

After identifying the topics included in the particular EHR and the possible connections between the patient and his/her relatives, the system needs to de-

cide which templates fit for the representation of the patient data. The possible templates are stored in a resource bank, which contains templates for common information as well as specific templates for the particular medical sub-domain. These templates are derived after study of representative amounts of EHR. A sample template is shown in Fig. 3. To narrow down the search to support the choice, the system uses a domain ontology. The chosen template is included in the graph node for the current patient's EHR. The system maintains four types of ontologies - of symptoms, of diagnosis, of drugs and a shallow ontology of body parts.

The Post Processing Module recognizes important NP and VP chunks using the lexicon and partial grammar rules. Some efforts are needed to determine the VP chunks due to the telegraphic style of the medical reports which rearly contain complete sentences.

The Extractor determines the patient's symptoms or diagnosis which are reported in the current EHR. The module for Filling Templates tries to fill in the information for each symptom, diagnosis and treatment which are foreseen by the chosen template.

The Relation Analysis Module identifies four types of relations: *is-a relations*; *Cause-Effect relations*; *Internal relations* between symptoms, diagnosis and treatment in one node of the patient (Fig. 4); and *External Relations* (Fig. 5) between the different nodes in the patients' chronicle graph as well as between nodes of the patient's relatives graphs.

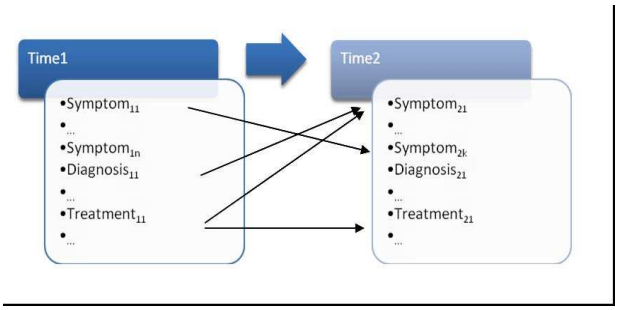

Fig. 4. Relations between the Patient's Chronicle Graph nodes

The Logical Form Generator creates conceptual graphs represented in first order logic, using the identified relations and the information which is already present in the templates (please note that some template slots might remain empty). The last step is to check again whether some empty slots in the templates can be filled in, given the context of all extracted information and the inference rules.

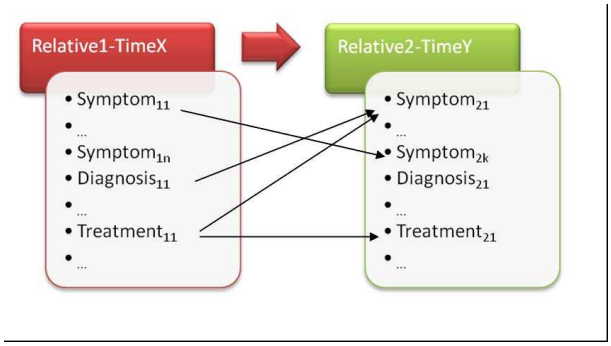

Fig. 5. Cause-Effect relations between template slots of two relatives

4 Relation Analysis Module

There are many kinds of relations between the concepts in a medical texts but we shall classify the relations in two general classes: *internal* (between the slots of one EHR) and *external* (between the slots of two different EHR, e.g. two records for family members).

Recognition of relations is crucial for proper text processing. For instance, the causal relation has significant importance in medicine, which deals with treatments and drugs that can affect or cure a disease. Due to this reason the causal relation is often explicitly indicated in EHRs using linguistic means (i.e. words such as *result, effect, cause* etc.). In some cases the specific phrasal structure helps to identifiy cue patterns, which work as indicators of the location of desired knowledge [15]. Unfortunately not all the cause-effect relations can be identified by keywords and phrasal patterns. There are more complex relations for which it is neseccary to process several discource sentences and to make inference in order to determine them. Khoo et al. [5] attempted to identify the location of causal relationship description using dependency subtree patterns. Very important task is to find effective cue patterns suitable for the domain and the mining goal. Usually the systems use cue patterns given a priori, presumably devised by domain experts for the prescribed tasks or collected by statistical information.

In our initial experiment we use about 150 EHRs in Bulgarian for diabetic patients. We investigated the specific verbalisations of the symptoms-diseases relations in the corpus. The selected cue expressions are ranked by frequency and include the most frequent adjectives, prepositions, adverbs and verbs: "оплаквания от" (complains) - 73% of its occurences in the texts signal for symptoms-diseases relations; "данни за" (there exists data for) - it appears at least 2 times per each EHR and 100% of the occurences denote symptoms-diseases relations; "поради" (because of) - 49.2% of the occurences in EHRs encode a symptom or a disease; "по повод на" (reason for) - 74.6% of its occurences in EHRs refer to symptoms and diseases, "съобщава" (inform) - 100% of its occurences signal symptoms-diseases relations (but this cue is tricky

because it appears mostly in combination with negation and it is not easy to identify the negation scope). All above mentioned cue phrases mark that the patient has some symptoms and diseases. Shallow ontologies for symtpoms, diagnosis and body parts supported the process of cue patterns extraction.

At present the Relation Analysis Module recognizes the following types of cause-effect relations:

- Patient: Status Relations
  - Between slots in one template - Symptom-Diagnose; Diagnose-Treatment
  - Between slots in two different templates - Diagnose-Symptom; Treatment-Symptom; Diagnose-Diagnose; Treatment - Treatment; Symptom - Symptom;
- Patient-Relative Status Relations
  - Cause-Effect relations between slots of two relatives - Symptom-Symptom; Diagnose-Symptom; Treatment-Symptom

We take into account three major types of cue patterns:

- Symptoms and conditions of diseases
- Verb expressions representing a relationship, interaction, or action
- Symptoms and conditions of diseases - for this type of patterns we use templates with predefined relations and empty slots for the concepts (symptoms, diseases), as well as slots for characteristics representing the condition.

Example for a diabetic patient: Постъпва за 1 път в клиниката, по повод на обща отпадналост, ацетонурия, високи стойности на кръвното налягане, а от няколко дни има повръщане. Заболяването е установено преди 4 години при измерване на кръвна захар, поради обрив на лицето. Въпреки назначеното лечение с манинил и диапрел няма подобрение.

This is the 1st visit of the patient to the clinic with complains of general weakness, acetonoria, high blood preasure, and sickness since few days. The disease was detected 4 years ago by the high blood sugar measurement, made because of a face rash. Despite of the treatment with medicamentations like Maninil and Diaprel there are no chaneges for better.

After analysis and chunking of the first sentence we obtain:
Постъпва{Постъпва.V+IPF+I:R3s:E2s:E3s} за{за.PREP} 1{gb} път {път.N+M:s} в{в.PREP} клиниката{клиника.N+F:sd}, по{по.PREP,по.PC} повод {повод.N+M:s} на{на.PREP} обща{общ.A+GR:sf} отпадналост {отпадналост.N+F:s} ацетонория{} високи{висок.A+GR:p} стойности {стойност.N+F:p} на{на.PREP} кръвното {кръвното .A:sn, кръвно.ADV+MNN} налягане {налягам.V+IPF+T:VNs, налягане.N+N:s}, а{а.CONJ} от{от.PREP} няколко {няколко.PRO+IDF:ms} дни {ден.N+M:p:c} има {имам.V+IPF+T:R3s:E2s:E3s} повръщане {повръщам.V+IPF+T:VNs, повръщане.N+N:s}

The Extractor uses a cue pattern (Fig. 6) for each symptom in order to process as much as possible words and phrases from the text and to send to the

Templates Filling module the necessary information.

To apply a pattern, there is a minimal requirement to fill in its obligatory slots, which in this case are only:

```
[HAVE]->(AGNT)->[PERSON]
      ->(THME)->[SYMPTOM]->(CHAR)->[CONCEPT]
```

The remaining slots are optional and they are filled in when additional infromation is present. The extractor generates the following CGs for the sample sentence 1:

```
[HAVE]->(AGNT)->[PERSON]
      ->(THME)->[SYMPTOM]->(CHAR)->[Weakness]->(ATTR)->[General]

[HAVE]->(AGNT)->[PERSON]
      ->(THME)->[SYMPTOM]->(CHAR)->[Acetonoria]

[HAVE]->(AGNT)->[PERSON]
      ->(THME)->[SYMPTOM]->(CHAR)->[Blood preasure]->(ATTR)->[High]

[HAVE]->(AGNT)->[PERSON]
      ->(THME)->[SYMPTOM]->(CHAR)->[Sickness]-
                              ->(ATTR)->[since few days ago]
```

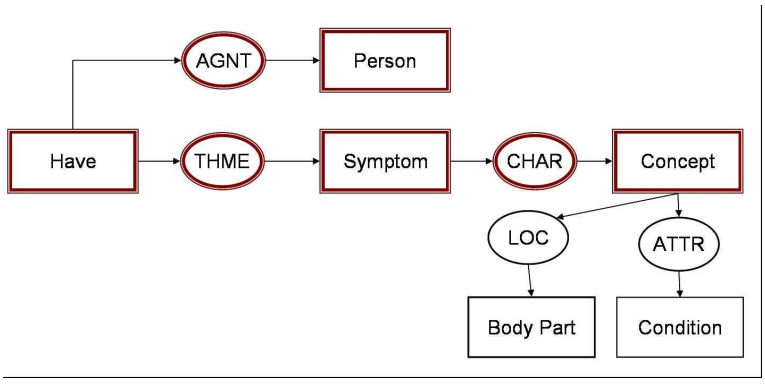

Fig. 6. Cue pattern for a symptom

These "elementary" cue patterns help us to fill in the templates by the words and phrases which are encountered in the text. Implicit relations are found in

this way - e.g. AGNT, THeME, CHAR, ATTR, LOC. They do not correspond to specific words in the EHR. Another type of cue patterns - "Verb expressions representing a relationship, interaction, or action" - support the discovery of relationships between the patients' template slots as well as relations among several slots in the patient's chronicle and his/her relative tempales.

Most generally, the relations between the slots in the different sections of one EHR connect each Symptom with the corresponding Diagnose and each Diagnose with the corresponding Treatment. To discover such relations we apply statistically collected cue phrases like *effect, results, influence, changes, achievement* etc. For instance, the cause-effect relation representing the result after the treatment "лечение - подобрение" (treatment - improvement) can be found in the sample sentence above: Въпреки назначеното лечение с манинил и диапрел няма подобрение. (*Despite of the treatment with medicamentations like Maninil and Diaprel there are no changes for better.*)

To discover relations between slots in different patient's nodes of the Patients' Chronicle graph we use the rich temporal information in the EHR. The most frequent cue phrases include time, dates, years, months, after, before, etc. The main task is to find the last and the first occurence of each symptom and diagnose and to connect them to the corresponding Treatment. The example shows the temporal relation "disease was detected 4 years ago": Заболяването е установено преди 4 години поради измерване на кръвна захар, поради обрив на лицето. (*The disease was detected 4 years ago by the high blood sugar measurement, because of a face rash.*)

Relations between slots of the patient's node and his relatives can be detected by searching for common symptoms, diagnosis and comparing the treatments. The more complex cue patterns, however, extract too many irrelevant phrases and the result needs human revision. On the other hand the too specific cue patterns generate only few results. The process of relations identification should be iterative in order to achieve more significant reesults.

5 Generation of CG in Logical Form

The CG generator collects all the information from the patient's node templates and the information about the corresponding relations between templates' slots. Here we sketch the CG generation algorithm:

- step 1: For each template T_i we construct one graph G_i using maximal join operation for the corresponding common concepts in the template slots.
- step 2: For each internal relation R^1, R^2, ..., R^k between slots in the template T_i - add consequently relations to the graph G_i between the corresponding concepts.
- step 3: Cluster the set of T_i in subsets depending on whether the templates are linked by external relations. A given template T_k belongs to a cluster C_m if and only if there exists template T_n from C_m and an external relation between T_k and T_n. The resulting clusters contain interlinked templates.

– step 4: For each external relation between slots in different templates T_i and T_j which belong to the same cluster - construct a relation between the corresponding concepts in G_i and G_j and generate a new graph G_{ij}'
– step 5: Join all new graphs G_{ij}' belonging to one cluster
– step 6: Represent all constructed graphs as Logical forms (LF).
– step 7: The presented EHR contains as many LFs as the number of clusters.

This algorithm ensures the production of connected conceptual structures, which encode interlinked information in the EHRs. Several issues have to be mentioned here.

The bottom elements in the construction are the system patterns - like the one at Fig. 6 - which shape the extracted words/concepts into conceptual structures. The pattern relations are either present in the text explicitly, or are introduced by default as thematic roles like CHAR, ATTR, AGNT, THeME. Joining the simple patterns at step 1, we obtain one conceptual graph:

```
[HAVE]->(AGNT)->[PERSON]
      ->(THME)->[SYMPTOM]-
                  ->(CHAR)->[Weakness]->(ATTR)->[General]
                  ->(CHAR)->[Acetonoria]
                  ->(CHAR)->[Blood preasure]->(ATTR)->[High]
                  ->(CHAR)->[Sickness]->(ATTR)->[since few days ago]
```

The internal relations between the templates enable joining of conceptual structures which correspond to separate sentences and paragraphs. The relations correspond to referential links between text fragments. We assume that there is no referential ambiguity since the domain is strictly fixed and the language is rather specific. Step 2 allows to connect conceptual structures that are linked because of linguistic evidences in the EHR text. For instance, if for the sample patient above it is mentioned in the same EHR paragraph that he/she was diagnosed by Diabetes, then in the same template we would find the following graph:

```
[HAVE]->(AGNT)->[PERSON]
      ->(THME)->[Disease]->(CHAR)->[Diabetes]
```

After steps 1 and 2 we would oobtain the following graph:

```
[HAVE]->(AGNT)->[PERSON]
      ->(THME)->[Disease]->(CHAR)->[Diabetes]
      ->(THME)->[SYMPTOM]-
                  ->(CHAR)->[Weakness]->(ATTR)->[General]
                  ->(CHAR)->[Acetonoria]
                  ->(CHAR)->[Blood preasure]->(ATTR)->[High]
                  ->(CHAR)->[Sickness]->(ATTR)->[since few days ago]
```

Steps 3 and 4 enable to build groups of interlinked templates and graphs in case of external links (reflecting links among family members). Steps 5 and 6 juxtapose one logical statement to conceptual structure which encodes connected

facts in the text. Here under 'connection' we mean again explicit linguistic references.

In general, the join operation may unify different unspecified instances of the same concept type, which is problematic from a knowledge representation perspective. However, studying EHRs we discover that often each word occurrence refers to one instance - e.g., the blood pressure of the patient. We are currently investigating this issue, in order to motivate our algorithms for construction and unification of conceptual structures. This on-going work will continue in the next year by tests and elaborations of our empirical approach which is tailored to the specific domain.

6 Conclusion and Future Work

The approach presented in the papers aims at domain conceptual modeling; it extracts facts from information described as unstructured text in natural language. Since language technologies operate on words and phrases, the atomic extracts are knowledge chunks corresponding to domain-specific templates. The suggested scenario applies the typical IE settings; for instance all words, which remain outside the templates, are considered as unimportant. Another issue, typical for IE, is the fact that the implicit relations are explicated due to the template slots. After the identification of template fillers, which are most often noun phrases, mining for patterns signaling relations is needed. At present we evaluate the precision and recall of the first experiments. It is obvious that the whole process of extraction and modeling will be iterative with several development cycles.

Conceptual graphs are well-suited to serve as primary patterns because they are adjusted to natural language applications. They also provide a well-defined join operation, assuming that graphs can be "merged" on their common concept instances. Our intuition and text examinations show that very often the words occurring in the EHR text point to single instances. This referential particularity is another important issue to be studied in the near future.

7 Acknowledgement

The research work presented in this paper is partly supported by grant №ID01/157 "Effective search of conceptual information with applications in medical informatics", funded by the Bulgarian National Science Fund in 2009-2011.

References

1. Boytcheva, Sv., 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.

2. Friedman C, Hripcsak G, Shablinsky I. An evaluation of natural language processing methodologies. In Chute CG, ed. Proceedings 1998 AMIA Annual Symposium. Phil: Hanley and Belfus, 1998, p. 855-859.
3. Friedman, C. and Hripcsak, G. Natural language processing and its future in medicine. Academic Medicine. 1999;74(8), p.890-895.
4. George Hripscak, Carol Friedman, Philip O. Alderson, William DuMouchel, Stephen B. Johnson, Paul D. Clayton. Unlocking clinical data from narrative reports: a study of natural language processing. Annals of Internal Medicine, 122:681-688, 1995
5. Christopher S. G. Khoo, Syin Chan, Yun Niu. Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. In Proceedings of 38th Annual Meeting of the ACL, Hong Kong, 2000
6. Leroy, G., Chen, H., and Martinez, J.D. A Shallow Parser Based on Closed-class Words to Capture Relations in Biomedical Text. Journal of Biomedical Informatics (JBI) vol. 36, pp 145-158, June 2003.
7. Lee, C.H., Khoo, C., and Na, J.C. (2004). Automatic identification of treatment relations for medical ontology learning: An exploratory study. In I.C. McIlwaine (Ed.), Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference (pp. 245-250). Wurzburg, Germany: Ergon Verlag.
8. Electronic Health Records Overview, National Institutes of Health and National Center for Research Resources, The MITRE Corporation, Center for Enterprise Modernization McLean, Virginia, USA, 2006.
9. MedLEE - A Medical Language Extraction and Encoding System, http://lucid.cpmc.columbia.edu/medlee/
10. Schulz, S., B. Suntisrivaraporn, and Fr. Baader. SNOMED CT's Problem List: Ontologists' and Logicians' Therapy Suggestions. In Proceedings of the Medinfo 2007 Congress, Studies in Health Technology and Informatics (SHTI-series).
11. Natural Language Processing in Medical Coding. White paper of Language and Computing (www.landcglobal.com). April 2004.
12. Neubauer, T. and B. Riedl. Improving Patients Privacy with Pseudonymization. In S.K. Andersen et al. (Eds.) eHealth Beyond the Horizon —Get IT There, IOS Press, pp. 691-696, 2008.
13. Gindl, S. Negation Detection in Automated Medical Applications, a Survey. Vienna University of Technology, Institute of Software Technology & Interactive Systems, Asgaard-TR-2006-1, October 2006.
14. Sowa, J. Conceptual Processing in Mind and Machines. Reading, MA, 1984.
15. Shimbo, M., S. Tamamori and Y. Matsumoto. Finding cue expressions for knowledge extraction from scientific text early results. In: Pacific Knowledge Acquisition Workship 2004, 09/08/2004 - 10/08/2004, Auckland, New Zealand, 2004