MedAT: MEDICAL RESOURCES ANNOTATION TOOL

M. Žáková*, O. Štěpánková* and T. Maříková**

*Czech Technical University/Department of Cybernetics, Prague, Czech Republic
** Charles University, 2nd Medical School, Institute of Biology and Medical Genetics, Czech Republic

zakovm1@fel.cvut.cz, step@labe.felk.cvut.cz, tatana.marikova@lfmotol.cuni.cz

Abstract: Medical Resources Annotation Tool supports creation of semantic annotations of medical records or similar semi-structured texts. In MedAT annotation is created by filling of forms. The forms are generated dynamically from an ontology, that is developed with medical doctors to describe the structure of the documents to be annotated. The facts obtained by the annotation are saved into a knowledge base in OWL or OCML ontology formalism. The facts are also saved into a relational database in a form that makes them suitable for relational data mining.

Introduction

Machine learning and data mining (DM) offer numerous semi-automated methods how to find characteristic patterns and how to gain better insight into data records stored in computer databases. It is tempting to apply these methods in medical domains to identify e.g. hidden relations among some symptoms, best predictors of certain diseases or possible interactions between medications. Records of patients collected in hospitals or in consulting rooms of medical specialists represent valuable source of data, which could serve well the purpose.

Unfortunately, DM cannot be applied directly to the medical records most often. One of the main reasons is the format of data stored in hospital archives – while its basic format is a text file written in a natural language, DM tools are ready to work with source data stored in a structured form, namely as a set of tables. In the source table for DM, each object (e.g. a patient) can be represented by a row in such a way that the values of its attributes appear in the columns of the corresponding name.

It is clear that any DM project concerning medical data has to be aware of this gap and design an appropriate transformation of the considered data in patient's text files into a new form, namely a database. This process is too demanding and time-consuming to be done manually for rich sources of data. On the other hand full automation of the process is still out of reach of state of the art in natural language processing. That is why we are designing an interactive tool which can support the process starting with the patient's medical files and leading to a relational database ready to be subjected to DM.

Case study

A typical target application of our system appears in a specialized medical consulting room e.g. genetic consulting room with up to few hundreds patients. The doctor keeps patient records with regard to a particular disease and he/she suggests examinations in different laboratories. Reports of these examinations and of treatment at other departments have to be incorporated into patients records. The doctor needs to be able to search the records without any specialized knowledge of database technologies. The records should be stored in format that is suitable for relational data mining [1].

Patients' records are semi-structured. They consist of structured data such as blood pressure level, red blood cells count etc. However a significant part of the report consists of notes made by the doctor. These notes are written in natural language, often in rather complex sentences. They also contain a lot of abbreviations, some of them are non-standard. The records are therefore very heterogeneous, e.g. due to the different examinations that the individual patients undergo. Since a significant part of reports is in free text, further heterogeneity is introduced by personality of the doctor. For several hundreds of reports general text mining algorithms do not give sufficiently good results [4].

A viable option is the use of a parsing algorithm and gazetteers [6]. However specialized gazetteers are not currently available in Czech, therefore one function of our system is to aid in creating a specialized gazetteer on basis of the available records.

System description

Medical Resources Annotation Tool (MedAT) supports creation of semantic annotations of medical records or similar structured texts. It enables the user to open a medical report in text format and to transfer the information contained in the report easily into dynamically generated forms which capture the basic structure of the report.

MedAT uses the framework of Dynamic Narrative Authoring Tool (DNA-t), which was developed for semi-automatic creation of semantic annotations of historical narratives using conceptual graphs [9] which were later used to explore concepts across the individual narratives [7].

The system has a flexible modular architecture. Its

ISSN: 1727-1983 © 2005 IFMBE

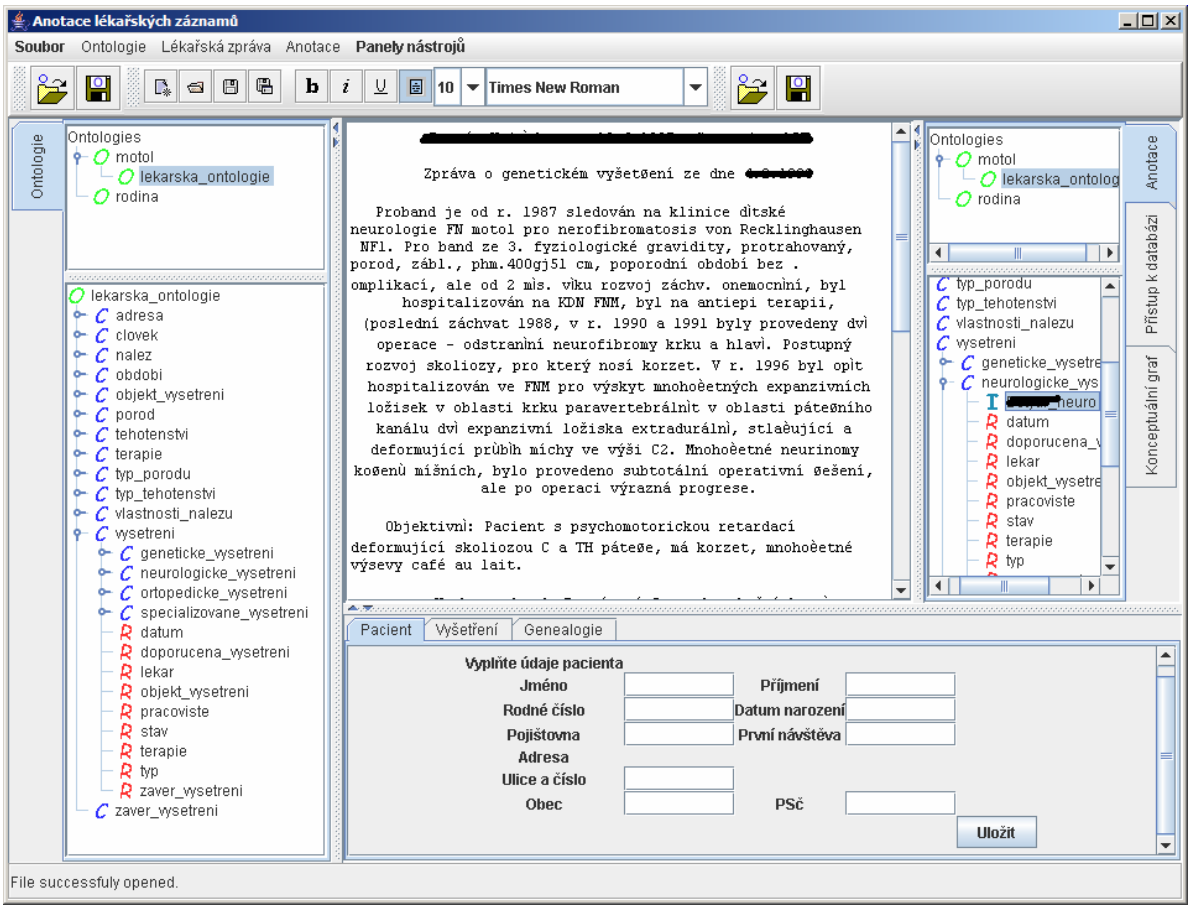

Figure 1: The GUI of MedAT – pilot version of the system is developed in Czech

core is formed by module for ontologies, module for displaying and annotation of documents, module for forms generation. In addition to the core modules, there has to be at least one concerning the output: namely, the module for saving information into relational database or the module for creation of a knowledge base. Additional modules for data access and visualization such as genealogy tree viewer or module for SQL queries can be easily added.

MedAT also currently suppports adding of instances to the domain ontologies based on concepts appearing in the annotated documents and thus creating a customized version of the ontology. Adding classes to the ontology or modifying its structure is currently not allowed, since MedAT does not include any tools for consistency checking and ontology merging.

Creating Annotations

In MedAT annotation is created by filling of forms. The forms are generated dynamically from an ontology, that was developed to capture the structure of the used medical records. Information is transferred to the forms by drag and drop. In case there is a set of possible values of a particular field available in form of instances of the class that corresponds to the given field, these values are filled into a dropdown list and the user can then select the required value from the dropdown list on the form. A schematic example of annotation creating can be seen on Figure 2.

A form corresponds to one major class in the ontology and at the same time to one master table in the database such as *Patient*, *Examination* etc. Forms that are currently being used are placed on individual tabs of the tabbed pane, so that the user can easily switch among them.

It is planned to add partial parsing of the documents based on the ontology describing the report structure and a gazetteer of the used concepts. As a result, some information on the forms would be filled in automatically.

To keep track of the abbreviations used in the documents and to prepare way for the use of semi-automatic parsing of the documents, the user can link the abbreviation used in a particular document to an instance in the ontology. The information linking the abbreviation to that particular instance is then saved to the ontology and the abbreviation is on the form with consent of the user automatically replaced by the corresponing full term.

Filling of forms was used as an intermediate step of the annotation instead of conceptual graphs, which were used in DNA-t. This solution fits better the medical personnel accustomed to this way of data visualization. Medical reports are also semi-structured texts, so the visualization using forms to some extent follows the

ISSN: 1727-1983 © 2005 IFMBE

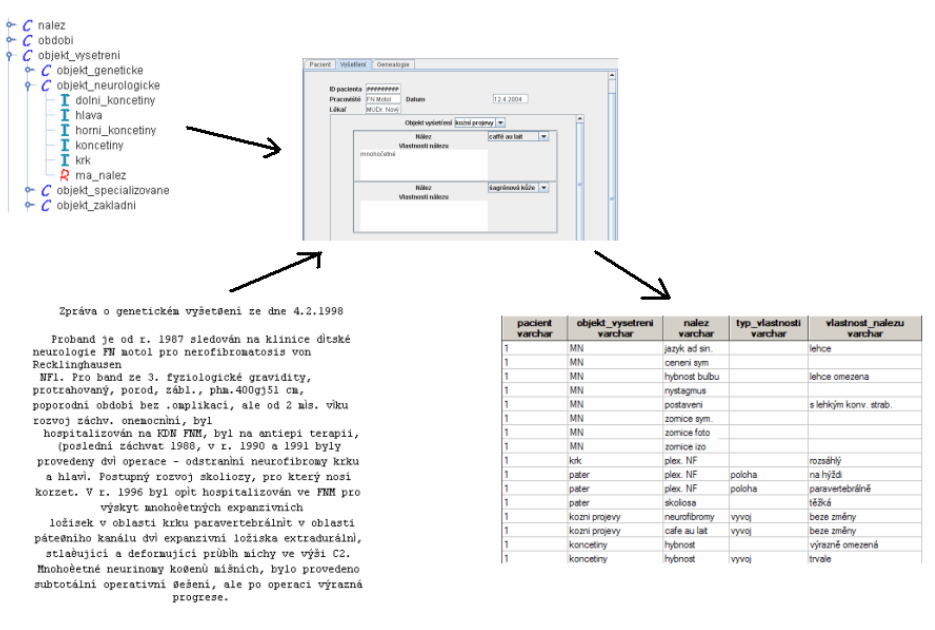

Figure 2: An example of annotation

structure of the paper reports. Also a higher degree of standardization is required for medical reports than was necessary for historical narratives for which DNA-t was primarily designed.

As the information is transferred from the document into the forms, the information is simultaneously automatically saved into the knowledge base that is based on the ontology that was used to generate the forms. The knowledge base is in a frame-based formalism of the Apollo ontology editor that is based on OCML. It can be automatically exported to Lisp and used for reasoning. Moreover, it can also be automatically exported to a set of logical formulas and facts that can be used directly for relational data mining.

Additional Functionalities

The knowledge base can be viewed as a tree. Some parts of the knowledge base can be exported and viewed in form of a conceptual graph. Some more specialized visualisations such as a tool for viewing genealogy tree in a form that is standard in medical community are currently being prepared.

MedAT also provides the user with the ability to explore the data stored in the database. There is a tab enabling the user to access the database directly. A list of predefined queries is maintained and gradually updated so that the user can only fill in the parameters. This is intended primarily for users with no knowledge of SQL. The users familiar with SQL can type whole SQL commands directly. The results of SQL queries are displayed in tabular form.

Relational database currently supported by MedAT data access module is PostgreSQL. However all SQL commands used in the application utilize only standard SQL features, therefore the application can easily be used with other database systems such as MySQL.

Ontologies

MedAT relies on ontologies on two levels: Task ontologies are used to describe the structure of the different medical records while Domain ontologies formalize knowledge about a specific domain e.g. ontology of family relations.

Task Ontologies

On the basis of procedures and structure of medical reports on examinations of various types recommended in literature and after consultation with a specialist an ontology was designed which attempts to describe the structure of the considered medical reports. In this case a hierarchy induced by the part-whole relationship. In the Apollo version of the ontology this hierarchy is captured only using slots with facets specifying properties that would apply in case of a proper part-whole relationship, such as cardinality = 1. In the OWL version of the ontology special properties *hasPart* and *partOf* have been designed as recommended in the W3C Working Draft [8].

There are two types of classes in the task ontology. Classes of the first type attempt to capture the elements that appear in the medical reports such as examination, finding, therapy etc. The attributes of these classes then represent properties or parts of the given element. For example examination has properties: date, doctor, laboratory, type and state. The report of an examination consists of these properties and also from elements: *object_of_examination, exam_conclusion, therapy and suggested_examinations*. These elements are also represented by classes and they are stored in the corresponding tables in the database. These tables are bound to the table Examination.

Classes of the second type are concepts, which describe a group of properties that appear in the reports

ISSN: 1727-1983 © 2005 IFMBE

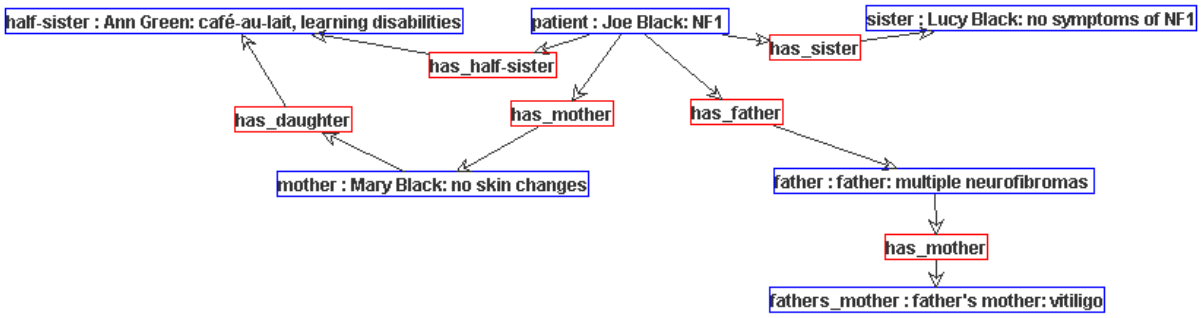

Figure 3: Visualization of a genealogy tree in a conceptual graph

e.g. properties of finding, type of disease. The individual properties are then represented by instances of the mentioned classes.

The task ontology serves as a basis for the structure of the generated forms. It also provides semantics of the data model of the relational database in which patients records are stored. There exists a one to one mapping between classes in the ontology and tables in the relational database. Slots correspond to attributes or relations between tables.

Further on, this ontology provides the structure of the knowledge base into which data about the individual patients and examinations are stored for the purposes of reasoning.

Therefore in case we are building a knowledge based system using an already existing database, the mapping between the ontology and the database schema can be used to fill the knowledge base automatically. On the other hand, the ontology could be used to generate a database schema for large and often changing systems.

The ontologies are available both in XML format in a frame-based formalism of the Apollo ontology editor [13] and in OWL-DL. Both formalisms can be easily automatically translated into logical formulae and facts that can then be used for data mining.

Use of Domain Ontologies

The second type of ontologies that is used in the application are domain ontologies describing concepts from a particular domain, relations between them and other background information about the domain expressed in form of logical formulae. It is anticipated that in future mostly third party ontologies available on the Internet such as GALEN [12], Gene Ontology [10, 11] and other ontologies describing diseases and medications will be used for this purpose.

Ontology of family relations was needed for annotation of reports about genetic examinations. Since there does not exist a freely available ontology that would correspond to the requirements of the application, an ontology of family relations was designed on the basis of the available reports about genetic examinations. It attempts to describe all family relations that could be relevant for the investigation of hereditary diseases. Every relationship is represented by a class in the ontology.

The relationships are described in more detail than is common for genealogies. There appear relations such as mother's mother, half-brother, mother's sister, descent in female line etc. These classes were created, because such terms are actually used in the medical reports and they also support some more advanced queries, such as choosing all relations from mother's side. Relations between the individual family relationships are defined using logical axioms. E.g. class grandfather is defined as an intersection of class grandparent and man. The logical axioms form an important and integral part of the ontology.

Two versions of the ontology were developed. The first version is written in OWL, which is becoming increasingly popular especially for sharing ontologies on the web. Family relations were defined mostly using OWL restrictions. However this version of ontology was not fully sufficient, since it is impossible to express relations such as half-brother in standard OWL.

A person Y is a half-brother of a person X, if Y and X have exactly one common parent and Y is a man. More formally this is stated in Formula 1.

$$\begin{aligned} &hasHalfBrother(x,y) = male(y) \wedge \\ &\exists a,b:(hasParent(x,q) \wedge hasParent(y,a)) \\ &\wedge \neg(hasParent(x,b) \wedge hasParent(y,b)) \end{aligned} \quad (1)$$

To express this relation, OWL rules would have to be used. However at present there exists no standard language for expressing OWL rules. Only some proposals can be found such as ORL [5]. Therefore the ontology of family relations was developed also using format of Apollo ontology editor. The formalism used in Apollo allows creating rules and exporting the ontology into OCML formalism [2], for which inference is defined.

The ontology of family relations also serves for visualization of genealogical tree of the patient's family in a conceptual graph to gain an overview of the occurrence of the disease in the family (see Fig. 3).
In general, domain ontologies should serve above all to aid standardization of the used terms and to speed up

ISSN: 1727-1983 © 2005 IFMBE

the annotation process by enabling the user to select e.g. family relationships from a drop-down list.

Results and Discussion

Our tool enables the user to open a medical report in text or XML format and easily transfer information contained in the medical report into dynamically generated forms, which follow structure of the report.

MedAT currently contains forms for basic records of personal data of the patients, information about genealogy and reports from specialized (e.g. neurological) examinations. Data from these forms are automatically saved into a relational database and they become ready for further analysis using e.g. SQL querries for generation of overviews and graphs or more sophisticated DM techniques.

Simultaneously, these data are saved as instances into an ontology, which describes structure of medical reports. This knowledge base can be displayed as a tree or as a conceptual graph, which is automatically generated from the annotation.

Conclusions

The tool is currently being tested at the Institute of Biology and Medical Genetics, 2nd Medical School, Charles University. It is planned that the resulting data will be analysed by methods described in [3]. The medical doctors working in the field of genetic counseling of neurofibromatosis type 1 (NF1) appreciate following features of the MedAT tool:

a) The process of creating the task ontologies helps the medical doctors to gain better understanding of the internal structure of the large set of gathered anamnestic data (like data on personal and family history, clinical and laboratory symptoms, specific mutations of NF1 gene, etc.) and follow-up studies. It is important that the system enables to modify the data structure dynamically, whenever appropriate.

b) The annotation tool enables to select sub-files (subsets of the registered patients) according to the content of particular symptoms (or their logical combinations) and carry out comparative mutation studies easily. These studies can enable to compare the main and supportive clinical criteria of NF1.

c) Thus, the characteristic symptoms of specific mutations can be easily correlated and, in such a way, the genotype/phenotype correlations can be detected. This will enable to develop highly efficient population genetic studies of NF1 for our population.

d) All the computer-supported work makes the administration of the collected data much easier as it enables to transform unstructured, on-the-fly written medical reports into a structured text. On the hand, in an ideal case the nearly complete reports (requiring just a minor "manual" finalization) can be retrieved from the structured text in the MedAT annotation tool.

Acknowledgements

The work has been supported by the grant 1ET101210513 "Relational ML for analysis of biomedical data" of the Czech nat. research program Information Society.

References

[1] WROBEL, S. (2001): 'Inductive Logic Programming for Knowledge Discovery in Databases' in DŽEROSKI, S., LAVRAČ, N. (Ed): 'Relational Data Mining', (Springer-Verlag, Heidelberg), pp. 74-84

[2] MOTTA, E. (1999): 'Reusable Components for Knowledge Modelling: Case Studies in Parametric Design Problem Solving', (IOS Press, Amsterdam)

[3] GAMBERGER, D., LAVRAC, N., ZELEZNY, F., TOLAR, J. (2004): 'Induction of comprehensible models for gene expression datasets by the subgroup discovery methodology', J. of Biomed. Informatics, Spec.Issue on BMI ML, **37/4,** pp. 269-284

[4] ANTOLÍK, J. (2005): 'Automatic Annotation of Medical Records' Proc. of the Medical Informatics Europe 2005, Geneva, in press.

[5] HORROCKS, I., PATEL-SCHNEIDER, P.F. (2004): 'A Proposal for an OWL Rules Language', International WWW Konference 2004, New York, USA, available online September 10

[6] LAVELLI, A., CALIFF, M. E. (2004) 'IE evaluation: Criticisms and recommendations', Proc. of the AAAI-04 Workshop on Adaptive Text Extraction and Mining 2004, San Jose, California, available online September 10

[7] MULHOLLAND, P., COLLINS, T., ZDRAHAL., Z. (2004): 'Story fountain: Intelligent support for story research and exploration', Proc. of the 9th In. Conf. on Intelligent User Interfaces 2004, Madeira, Portugal pp.62–69.

[8] RECTOR, A., WELTY, C. (2005): 'Simple part-whole relations in OWL Ontologies', URL: http://www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole/index.html, available online 10 September

[9] UHLÍŘ, J., KOUBA, Z., KŘEMEN, P. (2005): 'Graphical Interface to Semantic Annotations', Znalosti 2005 - sborník posterů, (Ostrava: VSB-TUO), pp. 129-132.

[10] THE GENE ONTOLOGY CONSORTIUM (2000): 'Gene Ontology: tool forn the unification of biology', Nature Genet, **25**, pp. 25-29

[11] GENE ONTOLOGY RESOURCES, Internet site address: http://www.geneontology.org/

[12] OPENGALEN Official Homepage, Internet site address: http://www.opengalen.org/

[13] Czech Technical University in Prague: APOLLO OFFICIAL HOMEPAGE, Internet site address: http://krizik.felk. cvut.cz/apolloch

ISSN: 1727-1983 © 2005 IFMBE