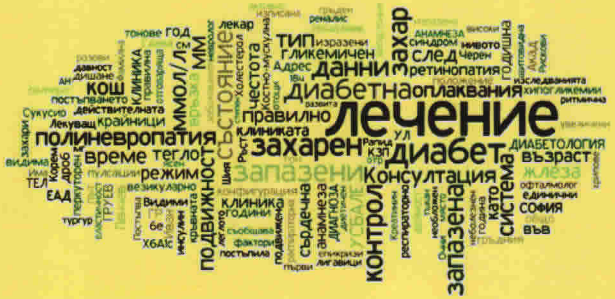
Effective Search of Conceptual Patterns with Applications in Medical Informatics

EVTIMA



National Science Fund
Competition Ideas 2008
Contract DO-02-292
2009 – 2011

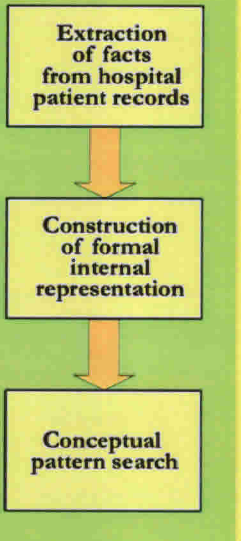
http://www.lml.bas.bg/evtima

Project Objectives

Participants

Results

Work plan


Extraction of facts from hospital patient records

Construction of formal internal representation

Conceptual pattern search

Effective Search of Conceptual Patterns with Applications in Medical Informatics

EVTIMA

National Science Fund
Competition Ideas 2008
Contract DO-02-292
2009 – 2011

http://www.lml.bas.bg/evtima

Actuality of the Task

The interest in automatic biomedical text processing is constantly growing worldwide. This is due to various reasons:

- *Exclusively fast increase of the amount of text documents in data bases of scientific publications, in the open Internet, as well as in the e-health area including patient-related documentation in electronic format;*
- *Increasing requirements to the quality of medical care;*
- *Ambition to improve the health management using automatic extraction of facts from medical documents etc.*

In this relatively new application domain, computational linguistics is facing a number of specific challenges:

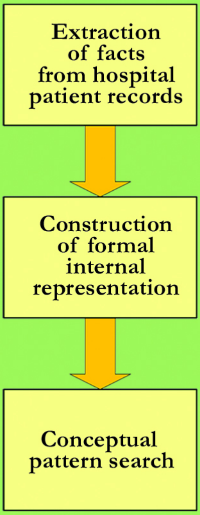
- *Very large terminological resources,* which hamper the development of domain dictionaries and electronic lexicons despite the natural limitations of the medical language style and genre;
- Necessity to *automatically interpret the text units as concepts and relations* from a predefined conceptual (semantic) domain model – but this declarative model has to be manually developed too;
- Lack of *standards for content and format of medical nomenclatures and classifications* as well as lack of *best practices for result testing, for reuse of software modules created to process different natural languages etc.*

Project Objectives

In a 3-years period, EVTIMA investigates the problems of effective pattern search by application of methods for fast discovery of repeating relationships among concepts in a very large archive of facts. The project uses an original idea for two-phase processing of conceptual information: at the first, preliminary off-line phase, all possible conceptual patterns in this domain are computed together with their generalisations and the whole resource is compressed in a minimal finite automaton with markers at the final states. In run-time, the conceptual pattern search is implemented as tracing the automaton paths, which is performed exclusively fast with linear complexity. The experimental tasks in the project aim at the development of a proof-of-concept demonstrator and the application of the idea, sketched above, to facts extracted from hospital patient records in Bulgarian language.

The objective of the first project phase is to settle the experiment basis, by developing of a software environment for automatic text processing of hospital patient records in Bulgarian. Moving towards these targets we have encountered other tasks of particular importance which enable the achievement of the main objectives and the continuation of the project activities beyond 2011:

- Development of linguistic resources and specific software modules for spell-checking of medical documents in Bulgarian.
- Design of an infrastructure supporting Bulgarian medical terminology with possible links to multilingual resources.
- Creation of annotated training corpora and structured test data sets which enable the development of language technologies oriented to medical text in Bulgarian.



Extraction of facts from hospital patient records

Construction of formal internal representation

Conceptual pattern search

Automatic Extraction of Facts from Biomedical Texts

The software systems for automatic knowledge discovery and information extraction usually perform several tasks in a pipe-line extraction procedure:

- Tokenisation: decomposition of the input text into words (symbol strings), alphanumeric literals, punctuation markers etc.
- Named Entity (NE) Recognition: finding the named entities and terms (e.g. person and organisation names, drugs), dates and quantities, important abbreviations and so on. The systems classify the strings, found in the texts, into NE groups. This procedure is supported by predefined lexicons with various named entities which are constructed using training corpora of representative texts.
- Identification and resolution of co-references between text units – e.g., an important term can be replaced by a pronoun or synonym.
- Recognition of relationships among the objects identified in the text. Often these relationships exist by default, since the document is written by a human expert addressing another human expert, who is familiar with the domain and the semantic links among the mentioned objects. The software system however "understands" the text by automatic calculation of the explicitly-named relations among the objects which are referred to in the text. Relationships explication is possible only by incorporation of predefined conceptual models of domain objects and relationships among them.
- Filling the entities and relationships, found in the text, into internal structures usually called event templates.

Extraction accuracy is measured by a sequence of tests on unknown documents. Usually the systems, performing knowledge discovery and information extraction, are tuned to find a predefined list of entities, relationships and facts of interest, without processing the whole input text.

Development of electronic resources in the medical domain – lexicons and declaration semantic models – is a real RTD challenge due to the large number of objects and relationships.

Participants

Two partners are involved in the EVTIMA project:

Linguistic Modelling Department (LMD) of the Institute for Parallel Processing, Bulgarian Academy of Sciences (IPP-BAS) – a leading Bulgarian RTD group in computational linguistics and language technologies, artificial intelligence and knowledge representation (http://www.lml.bas.bg). In its numerous international projects, LMD has acquired significant experience in the development of various natural language processing prototypes as well as linguistic resources in Bulgarian language.

Medical Informatics Department of the University Specialised Hospital for Active Treatment of Endocrinology (USHATE) "Acad. I. Penchev" at Medical University – Sofia. USHATE is a pioneer in the Bulgarian medical informatics and one of the founders of the Bulgarian medical network MedicalNet-BG. Its Hospital Information System stores more than 150 000 patient records at present, which comply the requirements of the Minimal Basic Data Set.

Experts in a number of related subjects are associated to the core EVTIMA partners: health managers, annotators of medical texts, and developers of intelligent systems. Two thirds of EVTIMA collaborators are doctoral students, post-docs and researchers younger than 40 years.

Work plan

During the first project phase the team is dealing with the following tasks:

- Studying, designing and detailed planning the project workflow and the final RTD result in order to meet both the user requirements and project objectives.
- Performing theoretical research investigations, related to the project topics.
- Developing the first version of software prototypes and evaluating their performance in experimental settings.
- Organising scientific events and participating in conferences and workshops in order to disseminate the project results.

The activities planned for the second project phase include more extensive research and expanding the experimental developments.

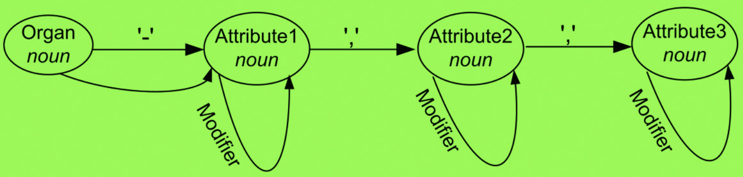
Automatic Processing of Hospital Patient Records in Bulgarian Language

Although structured in zones, the patient record texts are rather difficult narratives from the perspective of automatic text processing. They contain many numeric values and terms which are not included in the common Bulgarian dictionaries. The table below shows some statistics of the units occurring in the training corpus of hospital patient records. The significant amount of "unknown" words requires construction of a separate electronic lexicon of medical terminology.

	Strings	Non-language units		Unknown words		Recognised
.	Total	*Number*	%	*Number*	%	word forms
Training corpus	65 600	6 680	10,2%	13 370	20,3%	45 550

The texts contain Bulgarian and Latin terms in Cyrillic and Latin alphabet correspondingly; in addition, the Latin terms are often transliterated to Cyrillic. Some terms are recorded by various abbreviations. Complete sentences are rare, which hampers the application of traditional algorithms for syntax analysis. The narrative text consists prevailingly of declarative phrases, separated by punctuation marks. Considerations of important topics usually start in a new sentence but there is no practice to introduce new paragraphs in the various record zones (anamnesis - medical history, patient status, discussion). From the perspective of computational linguistics, these text features turn the patient-related narratives into very difficult input resource.

The extraction of patient-related facts in EVTIMA is performed using technologies for partial text analysis. Studying the training corpus, the team has developed linear rules for identification and recognition of terms and key phrases containing these terms. Practically we accept that the occurrence of a given language construction in the text is a positive statement about the presence of the uttered fact. There are also rules for treating various types of negative forms, which document absence or reduction of certain characteristics.



Graphically-defined morpho-syntactic rule for discovery of typical phrases in hospital patient records

Example:
Limbs – without oedema, preserved peripheral pulsations, onychomycosis.
organ attribute 1 attribute 2 attribute 3

Results

During the first project phase the team has developed various software prototypes for text processing of hospital patient records. The necessary computer lexicons and conceptual models, which describe the diabetes domain, are under construction. The terminological collections in Bulgarian language are compliant to the available Bulgarian versions of the nomenclatures ICD-9-CM and ICD-10. The conceptual resources play important role in text analysis because the explicitly-declared relationships among the concepts support the system's decisions regarding splitting the text into fragments discussing various subtopics. In this way the system "understands" how to decompose the free narrative text into separate structural segments. For the construction of conceptual models with Bulgarian vocabulary EVTIMA follows the practices and borrows examples from publicly available English resources in Internet, including the resources of UMLS (Unified Medical Language System) which are widely used in computational linguistics.

The software modules existing so far are integrated into a prototype system which extracts automatically facts concerning the patient diagnosis, the disease duration and the patient status. The extracted values are stored in a table, called template, which enables easy generation of internal formal and logically-consistent factual representation. In June 2010 the system is trained on a training corpus of 197 anonymised hospital patient records and the tests are run on a test corpus of over 1000 anonymised records. Extraction accuracy is evaluated using two typical characteristics:

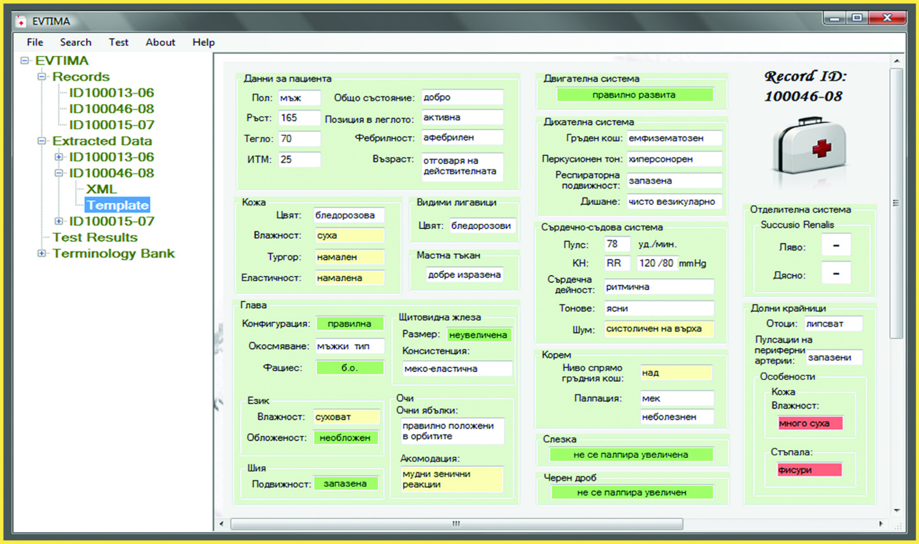
- ***Precision*** – percentage of correctly recognised entities among all extracted entities;
- ***Recall*** – percentage of correctly recognised entities among all entities available in the test corpus.

The procedures for text analysis are based on empirically-acquired rules of linear co-occurrence of terms in the patient record text. The more complex linguistic constructions require additional procedures for text processing. The current evaluation results are shown below. The assessment is done for the extraction of the diagnosis, disease duration and diabetic patient status.

	Diagnosis	Diabetes duration	Skin status	Neck status	Thyroid status	Limbs status
Precision	98,28%	96,00%	95,65%	95,65%	94,94%	93,41%
Recall	96,67%	83,33%	73,82%	88,00%	90,36%	85,00%

Results

Information extraction aims at the automatic identification of entities and relationships of interest and their classification as features describing the patient status. However free text descriptions contain a large variety of words, phrases and explanations; words might have various lexical forms and terms are embedded in phrases and paraphrases. The training corpus contains 93 characteristics only for describing the patient skin status. Correlations among attributes are automatically extracted from the training corpus. In addition to the text analysis and automatic extraction of status attributes, the current system prototype automatically classifies the words and phrases into patient conditions: good, fair, serious and critical. The system also contains knowledge about the default values which are used to fill in the empty slots in the structured patient status representation. Thus the prototype assigns values to the obligatory attributes and "normalises" the various characteristics turning them into items from a predefined scale of attributes. In the picture shown here the colours are meaningful: green fields denote default values, white ones correspond to good patient condition, yellow fields denote fair condition and the red ones signal serious and critical conditions.



Данни за пациента
Пол: мъж
Общо състояние: добро
Ръст: 165
Позиция в леглото: активна
Тегло: 70
Фебрилност: афебрилен
ИТМ: 25
Възраст: отговаря на действителната

Кожа
Цвят: бледорозова
Влажност: суха
Тургор: намален
Еластичност: намалена

Видими лигавици
Цвят: бледорозови

Мастна тъкан
добре изразена

Глава
Конфигурация: правилна
Окосмяване: мъжки тип
Фациес: б.о.

Щитовидна жлеза
Размер: неувеличена
Консистенция: меко-еластична

Език
Влажност: суховат
Обложеност: необложен

Очи
Очни ябълки: правилно положени в орбитите
Акомодация: мудни зенични реакции

Шия
Подвижност: запазена

Двигателна система
правилно развита

Дихателна система
Гръден кош: емфизематозен
Перкусионен тон: хиперсонорен
Респираторна подвижност: запазена
Дишане: чисто везикуларно

Сърдечно-съдова система
Пулс: 78 уд./мин.
КН: RR 120 /80 mmHg
Сърдечна дейност: ритмична
Тонове: ясни
Шум: систоличен на върха

Корем
Ниво спрямо гръдния кош: над
Палпация: мек
неболезнен

Слезка
не се палпира увеличена

Черен дроб
не се палпира увеличен

Record ID: 100046-08

Отделителна система
Succusio Renalis
Ляво: -
Дясно: -

Долни крайници
Отоци: липсват
Пулсации на периферни артерии: запазени
Особености
Кожа
Влажност: много суха
Стъпала: фисури

Structuring
patient status data
automatically extracted
from hospital patient records

Contact us:

Assoc. Prof. Galia Angelova, PhD, Dr. Sc., Bulgarian Academy of Sciences
http://www.lml.bas.bg/~galja
Assoc. Prof. Dimitar Tcharaktchiev, PhD, University Specilized Hospital for Active Treatment in Endocrinology
http://www.medicalnet-bg.org/dimitar

Published in June 2010