Knowledge Engineering Group

Information in Czech healthcare documents. Where is it hidden and how to extract it?

Karel Zvára
Ústav informatiky AV ČR, v.v.i. & EuroMISE s.r.o.
10th March, 2011

Presentation Structure

1. Source
   1. Archived documentation
   2. New documentation
2. Possible target structures and ontologies
   1. Code lists (SNOMED CT, LOINC, NČLP, …)
   2. Structures (EN13606, HL7 CDA R3)
3. Transformation
   1. Source text segmentation
   2. Exploiting iSpell
   3. Exploiting MeSH
   4. Transformation to target structure

Source: Archived Documentation

Documentation that has been gathered in the past. Usually, such documentation is available in the form of free text. Content of healthcare documnetation is outlined by ordinance No. 385/2006 Sb. (Ministry of Health). Different types of documents may have different structure.

Identification

Příjmení Jméno RRMMDD076
Praha 10 Ulice 5 211 2.11.2004

Brief overview

Nemocná s implantovaným kardiostimulátorem Biotronik Pikos v režimu VVIC pro FS od 25.5.2000 / celkem 54 měsíců /.

Examination result

Subj:bez zvl.obtíží
Obj.: spolupracuje, 81kg/166cm hlava, krk: bpn, TK: 125/70 , hrudník:kyfotický poklep plný, jasný, dýchání sklípkové,čisté, srdce:poklepově k mdcl.č.,akce srdeční pravidelná,na hrotě i na basi 2 temnější ozvy ,systolický šelest na hrotě s prop. do axily. břicho: měkké, palp.nebol., hepar, lien: nezvětšeny, tapottement bilat. neg. DK: bez otoků, lýtka nebol.
EKG:střídavě vlastní i stimulovaný rytmus komor, režim VVI
Parametry přístroje: frekvence základní: 69,8 /min.,
magnetická frekvence: 90,3 /min. interval : 860 ms
šíře impulsu: 0,5 ms
Funkce kardiostimulátoru : správná
Závěr.:

Conclusions

Implantace trvalého kardiostimulátoru v režimu VVIC pro fibrilaci síní
s pomalou odpovědí komor. Arteriální hypertense III.st dobře kontrolovaná. Mitrální rgurgitace II.s.t dilatace LS. Trikuspidální regurgitace. Arteriosklerosa povšechná.

Therapy

Therapie.:
Prestarium 4mg: 1/2 tb/d, HCHTH 1/2 tb/d, warfarinisace- ko INR, MonoMack 20mg: 1-1-0/d. Kontrola za 4 měsíce.

Source: New Documentation

Why? (Physicians may enter data to forms or even directly to EHR structures...)

Healthcare documents play different roles:

- … the basis for future medical care provision,
- … the basis for account rendering
- … is the body of evidence (mainly the defence of physician in case of prosecution]

As an evidence material created by medical care provider it should retain the feature of free expression. Care providers **demand this**.

Possible Target Structures / Ontologies

We need to know what we want to extract from documents. We just need to set the aim in the first place.

Possible structures:

- domain / purpose specific: e.g. „paramodel“ for Guideline Knowledge Representation Model (GLIKREM)
- standardized: EN 13606 / openEHR, HL7 CDA

Code books:

- according to structure used
- possibilities: SNOMED CT, LOINC, ICD10, MeSH, NČLP
- some may get translated by UMLS (NČLP is **not** mapped)

Source Text Segmentation

Input: monolitic unmarked/unformatted text file
Output: chain of basic containers

Tokenizer simply identifies creates a chain of containers. Each container contains either letters/digits or a single other character. Repeated occurences of single other characters are stored in single containers with repeat count held in container.

Example of tokenizer output: (for input: „Zvára Karel 760506/0001, 10.3.2011“)

1. raw: „Zvára“, count: 1
2. raw: „ “, count: 1
3. raw: „760506“, count: 1
4. raw: „/“, count: 1
5. raw: „0001“, count: 1
6. raw: „,“, count: 1
7. raw: „ “, count: 2
8. raw: „10“, count: 1
9. raw: „.“, count: 1
10. raw: „3“, count: 1
11. raw: „.“, count: 1
12. raw: „2011“, count: 1

Identifying Recognizers

Text split into individual containers is processed by „recognizers“, classes that recognize vocabulary words, formatted values and codelist values.

Basic recognizers:
SpecialCharRecognizer (produces TSpecialCharContainer)
– recognizes non-letter, non-numeric containers

NumberRecognizer (produces TNumberContainer)
- recognizes unsigned integer numbers (since +-,. characters are special chars)

DelimitedNumberRecognized (produces TDelimitedNumber)
- recognizes subchains of (regexp) number[\-delim\-number]+

URIRecognizer (produces TURIContainer)
- recognizes URIs in schemes *http*, *https*, *mailto* and *ftp*

Exploiting iSpell dictionary

GPL v2 License, good description of reasonably chosen flags

Not only spell-checker dictionary but also a good tool for PoS tagging. Unfortunately just half-way prepared.

Words are divided into several files. Named entities like city names, first names and Czech surnames are separated.

Inference flag combinations also hint word class. (see czech-grammar.aff).

ISpellRecognizer generates TdictionaryContainer
- contains word class identification

Implementation: all word forms generated to local Derby (JavaDB) database, select on lowercased field with index (1,25GiB database size)

Searching in Code Books

The biggest medical code system – SNOMED CT (Systematized Nomenclature in MEDicine, Clinical Terms) – has not been translated to Czech. It could be translated only if the Czech Republic became the national member of IHTSDO. This requires action by the government – the Ministry of Health – along with > 1M CZK yearly member fee.

The only medical codebooks available are MeSH (Medical Subject Headings) and set of code-books NČLP (Národní číselník laboratorních položek). NČLP also contains Czech version of ICD10 (International Classification of Diseases). On the other hand, MeSH is expensive for non-personal use (e.g. to be incorporated into software). The producer has to pay the price of 20.000 CZK (+10.000 CZK/year) and every customer further 2.500 CZK (+1.000 CZK/year) for just using Czech version of MeSH.

ICD10 and MeSH are both indexed by UMLS (Unified Medical Language System) Metathesaurus. This is vital for any professional use because standardized target structures (HL7 standards, EN13606/openEHR) use SNOMED CT and LOINC (which are also indexed by UMLS Metathesaurus.

Searching in Code Books (2)

Czech version of ICD10 is simply of no use for recognition.

- often uses abbreviations,
- abbreviations sometimes have different forms (even in the name of a single record, e.g. „Diabet.polyneuropat. při diab.“
- some terms are composed mostly of abbreviations: „J.deg.on.oč.víčka a periok.kr.“
- it does not contain unabbreviated version

Czech version of MeSH is much better

- cannot be used to identify diagnostic code (ICD10)
- not clinical-oriented as SNOMED CT and ICD10 (MeSH is a bibliographic code-book)

MeshRecognizer (produces TMedicalTermContainer)

- selects from local Derby (JavaDB) database (lowercased, indexed)
- uses subchains of containers, stop is translated to % (and then LIKE query is used)

Searching in Code Books (3)

What about SNOMED CT?
SNOMED CT is huge. And there is no Czech version. It also requires gathering license from IHTSDO (which I have obtained) to get complete database.

Although some medical terms are the same in Czech and in English, the usability is very limited. Only some terms (like „diabetes mellitus“) can be sucessufully identified just because of the same name/meaning.

Resolution
It looks like that the only usable code book is the Czech version of MeSH. But MeSH is non-clinical thesaurus created for indexing articles. It looks like there is no practically usable code book to be used for extracting information from medical documents.

Next Stage

As stated above, free expression is a required feature. New medical documents may have more structured information but (like with medical guidelines) the only authoritative version would probably remain in the form of free text.

Therefore the next task is creating a practically usable domain code book to be used at the cardiology domain (my sample documents are all from this domain).