Amenability of Czech Medical Reports to Information Extraction

Karel Zvára[1], **Vojtěch Svátek**[2]

[1] EuroMISE Centre, Institute of Hygiene and Epidemiology, First Faculty of Medicine, Charles University in Prague, Czech Republic

[2] University of Economics, Prague, Czech Republic

Abstract

Background: Patient's history, family history, diagnoses, medications and other information concerning patient's health and possible future treatment is usually incorporated in free-form narrative reports. Extracting relevant information helps giving the information to caretakers speaking other languages, utilizing modern techniques like reminding caretakers about conflicts with medical guidelines or collecting data for scientific use.
Objectives: The aim of this paper is to summarize the field of information extraction from free-form texts and to show results the author has achieved using simple methods for information extraction.
Methods: The lexical analysis and available Czech versions of medical codebooks were used in the first experiment.
Results: We show that narrative medical reports have a form so different from general texts and cannot be treated as general texts. Additionally available Czech codebooks were found insufficient to be used directly as dictionaries for term recognition.
Conclusions: New dictionaries of Czech medical terms need to be developed. Symbolic techniques have been found effective for recognition of pattern-specific values like Czech birth number or systolic/diastolic blood pressure values.

Keywords

Information extraction from texts, Czech medical reports, lexical analysis

Correspondence to:

Karel Zvára
EuroMISE Centre, Institute of Hygiene and Epidemiology,
First Faculty of Medicine of Charles University in Prague
Address: Katerinská 32, 121 08 Prague 2, CR
E–mail: zvara@euromise.com

EJBI 2012; 8(5):43–47
recieved: September 4, 2012
accepted: October 25, 2012
published: November 22, 2012

1 Introduction

The problem of transforming the text of medical records into structured form has been addressed by medical informatics research for decades. It is well known that the parsimonious writing style of the records, with frequent acronyms and abbreviations, as well as typos caused by time pressure, causes problems to state-of-the-art methods of information extraction from text. Even if partial successes were marked for English as a language with abundance of linguistic tools, nomenclatures, training corpora and, last but not least, stable word order [1], for many other languages the task remains extremely challenging.

The presented research focuses on medical reports written in the Czech language and influenced by the local legislation. The goal was to assess how much relevant information for subsequent transformation to structured form can be revealed via automatic analysis, using simple approaches to information extraction (i.e. those not relying on labelled training corpora).

In Section 2 we provide the taxonomy of information extraction methods, as broader context of our research (including methods planned for future work). In Section 3 we briefly characterize Czech medical records. Section 4 provides an overview of target nomenclatures (i.e. classes of information) and data structures (i.e. containers for information) to which the textual medical records (with special focus on Czech ones) should be converted so as to exhibit full machine-processability.

Section 5, eventually, deals with the application of information extraction on Czech medical records proper; after a brief overview of previous research we present our own research results, divided into three areas: part-of-speech analysis, specific pattern recognition and codebook mapping. Finally, Section 6 wraps up the paper.

2 Information Extraction Methods

Methods of information extraction may be divided into groups according to their subtasks [2], for example:

- Named entity extraction methods. The task of these methods is to find (and annotate) relevant textual properties like names, codes, dates, times, e-mail addresses.
- Co-reference analysis methods. The task of these methods is to find relations among individual words according to morphology of the input text (not specific pre-defined relations).
- Template filling methods. The task of these methods is to fill values found in text into pre-defined template. These methods may be used if there is known target structure (template) to be filled in according to input text.
- Relation extraction methods. These methods are used to extract pre-defined relations among extracted entities.

According to the type of extraction algorithm, information extraction tasks methods may be divided to two groups:

- Manual techniques are based on manually set rules, usually cascaded. This group includes techniques based on regular expressions.
- Trainable techniques are able to improve their ability to extract information from input automatically or under supervision. Trainable techniques usually need some supervision at least in the form of supplying annotations of input text. Trainable techniques include the bootstrapping technique (combining extracting with training). One of bootstrapping methods is "active learning" when annotating expert working with such a system annotates the document that the extraction method is least confident with.

Trainable techniques can be further divided into three groups:

- Symbolic techniques include e.g. Top-Down Induction of Decision Trees (TDIDT) – the "divide and conquer" algorithm (top-down approach) and "separate and conquer" algorithm (bottom-up approach).
- Probabilistic techniques include Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF).
- Other symbolic techniques include e.g. neural networks and support vector machines (SVM).

In the current paper we focused on named entity extraction using manual techniques. Applicability of such methods, in small scale, are a pre-requisite for using automatic techniques and addressing more complex extraction tasks in larger scale.

3 Czech Medical Reports

Czech medical reports are usually narrative reports (free-formed texts) formatted only by spaces, tabs and new lines.

The structure and even the obligation to create and keep medical reports has been incorporated into Czech legislative in 2001 [3] and [4]. The law set requirements for medical reports concerning the content and its form, especially structure.

Czech medical reports are therefore clinical texts with standardized structure. Common form and vocabulary is also determined by common education of physicians, their membership in professional organizations and their own interest on keeping credible documentation not only to enable long term care of the patient but also to defend themselves in judicial affairs.

3.1 Creation of New Medical Reports

New medical reports are usually created from templates and by copying and modifying last report. The reason for creating new reports by copying and modifying last report is economical. Doing "cut and paste" is fast and the physician will not forget to include mandatory information that does not change much during the time - like diagnoses, family history etc. This could lead to serious problems like neglecting changes in diagnoses. Similar problems have been observed also in other countries [5].

3.2 Content from External Systems

Some information comes from external systems in a form that can be simply copied, especially laboratory results. In the case of biochemical laboratory results, rows usually represent individual measurements and columns represent various properties like name of the measured variable, measured value, lower and upper limit. Sometimes simple graphics (created using symbols) is also provided.

3.3 Other Problems

Czech medical reports contain lots of typing error and abbreviations. That is not typical only for Czech medical reports. Individual abbreviations are usually not unambiguous, context is usually needed to decode correctly to find correct meaning. This problem has been also addressed by other authors, see [6].

Table 1: Parts of speech found in narrative reports (total annotations average).

	Annotations (avg per report)	Avg annotations: tot. tokens
Noun	75	30,32 %
Adjective	23	9,3 %
Pronoun	0	0 %
Number (non-digits)	0	0 %
Verb	17	6,87 %
Adverb	3	1,21 %
Preposition	0	0 %
Conjunction	0	0 %
Particle	0	0 %
Interjection	0	0 %

4 Target Structures and Nomenclatures

4.1 Nomenclatures

Target nomenclatures need to by recognized by users and/or their tools (information systems). Internationally and nationally (in the Czech Republic) the ICD (International Classification of Diseases) nomenclature is recognized and commonly used in medical reports. Concerning laboratory reports, SI units are also widely internationally used.

Concerning other codes, international and Czech national use differ greatly. In the Czech Republic, clinical information systems widely use the National Codebook for Laboratory (NČLP, Národní číselník laboratorních položek) which is not one codebook but tens of codebooks, some of them derived or copied from other codebooks (contains e.g. Czech version of International classification of diseases - ICD10).

Internationally, there exist more or less complex nomenclatures, specifically IHTSDO's SNOMED CT (Systematized Nomenclature in Medicine Clinical Terms), Regenstrief Institute's LOINC (Logical Observation Items Names and Codes) and Health Level Seven's Vocabulary.

These internationally recognized nomenclatures are administered by some legal entity and indexed by the National Library of Medicine and its UMLS (Unified Medical Language System). UMLS indexes more than 100 code-books and maps individual coded items to its own concepts, while maintaining network of relations between individual concepts. This way more-or-less accurate mapping among different nomenclatures is made possible.

In addition to UMLS, some nomenclature maintainers are trying to further formalize their nomenclatures to specify an ontology of described field. There are also initiatives that are trying to join partially developed ontology parts into complete ontologies (like OBO Foundry that is aimed at biomedical and biochemical ontologies).

4.2 Structures

Medical reports are non-formalized status documents describing patient's current status, observations and decisions/actions made. There are several influential organizations that concern themselves with formalizing electronic clinical documents, specifically TC 251 of CEN, Health Level Seven, ASTM American Society for for Testing and Materials) and openEHR Foundation.

Health Level Seven develops the CDA (Clinical Document Architecture) specification. It is designed to formalize administrative information, to annotate medical report on the level of report parts but allows to formalize individual clinical observations. Health Level Seven standards are usually developed using top-down approach (from general to specialized), the development is slow but the result is usually robust.

ASTM developed the Continuity of Care Record (CCR) standard. It represents just current state of the patient, so it is a state-report. Being developed not from the top but according to requests from users, CCR is more practical but less robust than CDA. ASTM and Health Level Seven together developed technical implementation of CCR using CDA. The result (CDA document containing CCR) is called Continuity of Care Document (CCD).

From the European perspective, the most important standardization of formalized electronic health record came from CEN, the EU normalization institution. CEN developed EN 13606 which has been adopted also by ISO. EN 13606 is usually referred to as "EHRcom". EHRcom specifies general way to formalize information commonly found in medical reports. It uses SNOMED CT, LOINC and other internationally used classification systems.

There are also projects which aim to standardize (minimal) content of electronic health record.

The epSOS[1] projects concerns also with a kind of minimal electronic documentation needed for urgent care of the patient. epSOS has published a specification of Patient Summary (PS) which is also mapped to existing european EHR standard EHRcom (EN 13606).

[1] European Patients Smart Open Services

Table 2: Delimited numbers recognition results.

	Found	Min. found	Max. found	Average
Blood pressure: SBP/DBP	434	0	12	1,62
Personal identification number	77	0	1	0,29
Not identified	268	0	6	1

5 Automated Analysis of Czech Medical Reports

First studies on automated (information-extraction-based) analysis of Czech medical reports were published in [7] and [8].

In the study [8] the regular analysis was used for information extraction. The paper [7] concluded that lexical analysis cannot be used because Czech medical reports are usually not made from whole sentences and the punctuation is almost not used.

The study [8] continued in the study published in [7] and enhanced regular analysis with some linguistic analysis. There were not used any codebooks and slightly better results were achieved in [8] than in [7].

We have studied the possibility of lexical analysis, recognizing specific patterns (like Czech personal identifiers or systolic/diastolic blood prossure) and using available code-books before. Partial results where published in [9].

5.1 Lexical (Part-of-Speech) Analysis

In order to analyze the distribution of differents parts of speech in the records, we reused the Czech iSpell dictionary from Petr Kolář, which was originally designed for spell-checking. The original version can be used for part-of-speech (PoS) tagging with just minor additions. Further, more complicated, addition would allow detection of inflection and gender but that has not been done because of poor results achieved from PoS tagging.

The Czech iSpell dictionary contains 260.679 basic words expanded to 4.624.350 words (some with exactly same expression but with different gender or part-of-speech tag) using affix rules. High number of annotations is determined by multiple annotations of recognized words.

Processing 268 narrative reports with a total of 66.286 tokens gave the results shown in Table 1.

5.2 Recognizing Specific Patterns

A relatively easy (though not trivial) task for information extraction consists in recognition of sequences of numerals with specific meaning. We focused on two common types of information, blood pressure values and the personal identification number of the patient. Specific combined numeric patterns were recognized with symbolic rule-based methods (similar to regular expressions). Different meanings were distinguished by fixed rules. In the case of blood pressure it was meaningfulness range of values, relation between parts of a pattern. In the case of personal identification number, the test of syntax correctness (lengths of parts) and meaningfulness has been used (personal identification number contains information on date of birth, gender and office that has allocated the number).

The Table 2 shows results of delimited numbers recognition.

There were no identified recalls, mostly because the rules for recognizing blood pressure and personal identification numbers were defined as strict. Both recognizers were defined for two decimals separated by slash with these properties:

- **blood pressure:** first number is greater than second, both numbers are positive, first number is lower than 500;
- **patient identificator:** rules for validation check of Czech personal identifiers were used (valid date and sex coded in the first number, identifiers corresponding to dates newer than January 1st 1954 are also checked for checksum).

5.3 Using Available Code-Books

Results of recognizing code-book terms have been published in [9]. Recognition of SNOMED CT and ICD10 terms has been totally unsuccessful. In case of SNOMED CT it has been expected because Czech version of SNOMED CT is not available. ICD10 has been used in the Czech version (part of NČLP code-books) but has been totally unsuccessful partly because specific expressed diagnoses have already been coded with ICD10 and partly because Czech names/descriptions of ICD10 terms are log and contain a lot of abbreviations.

The only successful coding system has been MeSH[2] in the Czech version. Even in the case of MeSH, we were able to recognize less than two terms per narrative report in average.

6 Conclusions

We can briefly summarize the main findings related to the three types of text analysis employed.

The **lexical analysis** is not a solution to information extraction from narrative reports written in Czech. The

[2]Medical Subject Headings

main reason is that narrative reports written in Czech are not regular sentences. This is manifested by the distribution of parts of speech, which clearly deviates from the distribution in contiguous text.

The main lesson learned from the lexical analysis part is that attention must be paid to typing errors and abbreviations. Both tasks should be solved alongside text extraction because abbreviations and typing errors are very often ambiguous. Therefore their translation to correct form needs context from other parts of the narrative report.

Symbolic techniques like rule-based filters or recognizing agents are good tool to recognize some specific numeric values. Such techniques can be effectively used to recognize blood pressure values and patient identification.

Looking up from standard **code-books** seems inefficient since most complete clinical code-books (especially SNOMED CT) are not available in the Czech language. Therefore some other code-book must be found, created or existing code-book translated.

Acknowledgements

This work has been supported by the specific research project no. 264513 "Semantic Interoperability in Biomedicine and Health Care", Charles University in Prague.

References

[1] Garcia-Remesal M., Maojo V., Billhardt H., Crespo J., Integration of Relational and Textual Biomedical Sources, Methods Inf Med, 2010

[2] Labský M., PhD thesis: Information Extraction from Websites Using Extraction Ontologies, Vysoká škola ekonomická v Praze, Praha, 2009 (Czech)

[3] Žďárek R., Vedení zdravotnické dokumentace a její náležitosti, Zdravotnické noviny, 3.6.2009 (Czech)

[4] Dostál O., Šárek M., Support for Electronic Health Records in Czech Law, European Journal for Biomedical Informatics, 2012

[5] Hammond K., Helbig S., Benson C., Brathwaite-Sketoe B., Are Electronic Medical Records Trustworthy? Observations on Copying, Pasting and Duplication. AMIA Annual Symposium Proceedings, 2003; 269-273

[6] Tsung O. Cheng, Letters to Editor; in: Medical abbreviations in Journal of the Royal Society of Medicine, Volume 97, 2004

[7] Semecký J., Zvárová J.(školitelka), Multimediální elektronický záznam o nemocném v kardiologii, Matematicko-fyzikální fakulta UK, Praha, 2001 (Czech)

[8] Smatana P., Paralič J. (školitel), Spracovanie lekárskych správ pre účely analýzy a dolovania v textoch, Technická univerzita v Košiciach, Košice, 2005 (Czech)

[9] Zvára K., Kašpar V., Identifikace jednotek a dalších termínů v českých lékařských zprávách, European Journal for Biomedical Informatics, 2010 (Czech)

