HOW TO RETRIEVE KNOWLEDGE FROM MEDICAL TEXTS EFFECTIVELY

P. Kolesa*,**, J. Antolík* and J. Ďurovec*

*EuroMISE Centre, Institute of Computer Science AS CR, Department of the Medical Informatics, Prague, Czech Republic
**EuroMISE s. r. o., Prague, Czech Republic

{kolesa, antolik, durovec}@euromise.cz

Abstract: With respect to previous projects ran at the EuroMISE, an interesting problem arises: how to transfer knowledge from medical texts written in a free text form into a structured format that is computer-interpretable. In the past, the problem of extracting patient data from medical record was solved by writing extraction rules (regular expressions) for every element of information that is to be extracted from the document. However, in general such approach is very time consuming and requires supervision of a skilled programmer whenever the target area of medicine or the text corpus are changed. In this article we explore the possibility to mechanize this process by automatically generating the extraction rules from a pre-annotated corpus of medical texts.

Introduction

One of the greatest challenges medical informatics faces nowadays is the representation of medical knowledge. Majority of the knowledge is represented in the human-understandable form – a form not suitable for processing by computer systems. What computer-interpretable form actually means depends on information represented. For example in case of electronic health record it means structured data. Ontologies and knowledge bases on the other hand are suitable representation for knowledge needed by decision support systems.

Transforming knowledge from human-readable to computer-interpretable from is a difficult task requiring lot of manual work by experts. On the other hand human-readable documents can be easily generated from computer-interpretable ones. In spite of this fact, only tiny part of medical knowledge is available in computer-interpretable form.

There are three reasons for this. First, medical knowledge is shared among medical experts, to whom the computer-interpretable representation is very unnatural. Further, medical experts often complain that representing knowledge in computer-interpretable form is too rigid and does not allow them to describe the reality. Finally, when physicians are forced to provide data in structured form, it takes them often more time than write them in free text.

Therefore, there is a great demand for a tool providing automatic knowledge extraction form medical texts. With respect to this problem, we explore two different areas.

The first project is transforming knowledge from a medical record written in a free text form into a structured electronic format represented by the Electronic Health Record (EHR). The second project is a knowledge extraction from medical texts.

Presently, medical records are usually written by a physician in a free text form. A straightforward solution of the information extraction problem would be to write a separate extraction rule for every single element of information that we are concerned with. However, such solution has some obvious disadvantages. First of all writing of these rules might be very time consuming, especially if we consider the fact, that the number of entries we need to collect can reach hundreds only within the scope of single medical area such as cardiology. Another problem that is sometimes overlooked is the fact that the process of creation of the extraction rules requires a close cooperation of the programmer and specialists within the target medical area. Such cooperation, however, might often turn out problematic and slow down the development even further. The final observation is that whenever the list of collected data elements is changed we have to repeat the whole process again.

The second project, running at EuroMISE, is concerned with transformation of existing medical knowledge from a free text to the computer interpretable form. In general this task is difficult and must be done manually without significant help of computer systems. But there exist collections of texts that can gain from computer support. Drug information leaflets (information about a drug usage and composition) are example of such a collection. A goal of this project was to develop a tool that will help to build a knowledgebase about drug-to-drug interaction and drug contraindications.

Transforming of medical knowledge into computer-interpretable form is a complex process that consists of several steps. For purposes of this paper let us consider only the very early stage: during that stage nominal phrases containing the requested knowledge are identified. In further stages those phrases are processed instead of the original text. We believe that extraction of such phrases can be done automatically.

ISSN: 1727-1983

Both mentioned projects require very high precision and completeness of the acquired data. That means that we have to find all the relevant information contained in the document and identify them correctly.

As a consequence of the previous facts, we would like to automate the process of generation of the extraction rules and thus eliminate the involvement of programmer from the process. Naturally, specialists from the target area are still required, because we will always need some source of information from which the extraction rules can be deduced.

In this article we will discuss our effort to build such automated system. Finally, let us note that both our projects focus on the Czech language, which is a natural consequence of the fact that it is the language in which we are capable to collect the pre-annotated medical texts. However, since many problems we came across so far are beyond the language barriers, we believe that the knowledge collected in this project will be applicable also to systems working with different languages.

Materials and Methods

Information extraction (IE) is a field of computer science that studies automatic extraction of information from textual sources usually written in natural language. As the description of this discipline indicates, it is of great interest for us since it may offer techniques that can help us solve the problem identified in the previous section. In fact several algorithms for automatic generation of extraction rules have been developed in this field such as RAPIER [1], SRV [2], WHISK [3], $(LP)^2$ [4], etc. In such systems the background knowledge comes in the form of pre-annotated corpus of texts. An annotation includes data about the start and end of subtext in the document and the semantic type of information that is stored in this subtext (e.g. pulse rate). Such corpus is then fed to the system, which after several iterations over the corpus offers the learned extraction rules as the output, where each of these extraction rules corresponds to a single type of information that is to be extracted.

The core of any system transforming documents written in natural language into structured electronic format is the information extraction algorithm. Instead of developing our own, we have decided to relay on a state-of-the-art project in this area of research. Particularly, we have decided to use the AMILCARE [5] system developed by Fabio Ciravegna at the University of Sheffield. AMILCARE is based on the $(LP)^2$ algorithm, which is a supervised algorithm that falls into a class of Wrapper Induction Systems using LazyNLP [6]. In the rest of this section we will briefly discuss the facts that supported our decision to use this particular system.

Probably the most important argument for integrating the AMILCARE system into our project is the performance of $(LP)^2$ algorithm compared to his other counterparts [5]. Another reason is that the AMILCARE system provides us with several means allowing us to supply it with some additional knowledge. For example gazetteers can be inserted into the system (for example a list of pharmaceutics or a list of possible diagnosis) or the input text can be extended with various tags that may help in the learning process.

Another advantage is that the NLP pre-processing phase is separated from the main system and thus allows us to use custom tools, which is especially important with respect to the fact that we will work with texts written in Czech language. Finally AMILCARE contains also a Java API, which enables us to easily integrate the extraction rules produced by the system into our own application. Generally, to our best knowledge, AMILCARE is currently the most mature system for automatic extraction rule generation, ready to be deployed in real world problems.

Within the scope of the discussed projects, several particular steps were conducted. First of all, we have created two independent corpora: one of medical records containing approximately 1000 health records, enriched with more than 140 different types of medical annotations and the second corpus consisting of drug information leaflets, containing more than 300 documents. In the second corpus only 3 types of annotations where used.

All annotations are provided by annotators, who process each document of corpus manually. Such annotations are necessary only for inducing and testing extraction rules. In the process of information extraction only created extraction rules are used and no manual work is necessary but supervising.

The process of preparation of the training corpus for the AMILCARE system continues in following way: every text has to be enriched with NLP information in order to provide additional tokens that can be exploited in the formation of the extraction rules, thus possibly increasing the performance of the final system. For this purpose we use the Prague Dependency Treebank toolkit [7] developed by Jan Hajič at the Charles University in Prague. This tool is capable of extracting various linguistic information from a document. We use primarily the lemmatization engine and also word tagging providing full part of speech information. Further we add some simple flags such as *capitalized* word etc.

The final step of the pre-processing phase is the integration of the NLP information, the annotations and the original texts into a single large file that serves as the input to the AMILCARE system. All the above described steps are mechanized by means of several interleaving PERL scripts. This way we have built a tool that receives the list of all original texts and annotations data as the input and outputs a single large file that serves as the input to the AMILCARE system.

Generally straightforward application of IE systems to biomedical data is problematic. The main reason comes from the fact that the performance of most IE systems depends on the performance of Natural Language Processing (NLP) tools that do the pre-processing of the input text and supply the IE system

ISSN: 1727-1983

with additional information. Since these NLP tools are usually developed and tested over corpuses of more

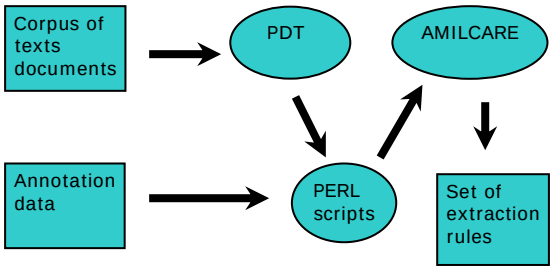

Figure 1: Corpus processing with a system using NLP and IE tools.

common types of texts such as newspaper articles , they do not achieve full performance when applied to biomedical texts [8]. This fallback in performance naturally propagates further in the given IE system. Although there already are promising results in retraining some NLP tools over biomedical corpuses it is unlikely that such applications will soon be available for less widespread languages such as Czech, because the costs of collecting appropriate annotated training data are high.

Unfortunately, even further problems arise when we focus our attention to the medical health records. The most important difference between them and the majority of other analysed biomedical documents is that medical records are not written for the purpose of publication. Consequently the overall quality of the captured text is significantly lower. One of the most important negative effects is that medical health records are usually poorly structured. Further they contain a large number of spelling errors that are very hard to detect by the IE tools. The use of character '*O*' instead of '*0*' or '*l*' instead of '*1*' is just another example of inconsistencies commonly contained in medical records. The virtual non-existence of division of the text into proper sentences prevents the application of more advanced NLP techniques such as part of speech tagging, which normally supply the IE system with valuable information. One of the most important goals of our projects is to find techniques that will overcome these problems specific to the medical health records.

On the other hand drug leaflets are common free texts. Thus information about document structure such as paragraphs and sentences can be used in IE. Also typing errors are quite rare in drug leaflets.

Results

For the medical record corpus the outputs of the learning process show promising results with respect to recall and precision. The data have also revealed several interesting insights into the textual data produced by physicians. The system also proved a suitable feature of being strongly biased towards precision. Some results can be seen in Table 1.

In case of the drug leaflet corpus the results are worse than for medical records. Precision and recall are lower by 44 and 16 per cent respectively. Such low rates are achieved although full part of speech tags and gazetteers are used. Results of the drug leaflet corpus are in the Table 2. The weighted results for both corpora are shown in Table 3.

Table 1: An excerpt of a table containing the results for medical record annotations. Results are in percents and are rounded up to integers.

TAG	Precision	Recall	F-measure
admin.birthnum	100	90	94
admin.contact.phone	95	97	96
admin.helinssoc	98	85	91
lab_exam.3glycerol	100	80	88
lab_exam.glycemia	100	77	86
lab_exam.cholest.hdl	97	80	88
lab_exam.cholest.total	97	79	87
lab_exam.thyr	100	77	88
lab_exam.uric_acid	100	81	89
operwhen	100	82	89
phys_exam.heartsound	100	80	88
phys_exam.height	100	81	88
phys_exam.pulse	100	75	86
phys_exam.thyr	100	80	88
phys_exam.weight	100	80	88
recomm.drugname	100	91	95
rf.dmtreat	100	77	87
rf.dmwhen	100	81	90
sh.allergy.todrug	100	83	92

Let us present some interesting observations. The first important fact is that out of the more than 140 different types of medical data that were selected for extraction only about half had occurrence higher than 20 in the approximately 1000 medical records included in the testing corpus. This means that we can build a system that may substantially reduce the amount of work physician have to conduct by looking only at relatively small subset of the collected data elements. In the rest of this section let us discuss only the more frequently occurring types of extracted information. This group subdivides into two subgroups – those cases where the performance of recall and precision reaches reasonable values and those which appear to be problematic (F-measure below 10). It is interesting that there are very few cases between these two groups – those with the F-measure in the range [10-50]. Another encouraging fact is that those cases which have reasonable values of F-measure thus are potentially useable already after this first preliminary training reach very high values of precession (but not necessarily also recall). This is a favourable condition, because it is usually acceptable for users when only partial help is provided, but it is frustrating to correct additional mistakes produced by the software. Table 1 contains the list of more frequently occurring annotations with performance statistics.

For medical record corpus we use a drug gazetteer to help to identify prescribed drugs. This improves the

ISSN: 1727-1983

drug slot recall approximately by 6 per cent while the precision remains the same.

Table 2: Statistics of slots searched in drug leaflet corpus. Results are rounded up to integers.

TAG	Precision	Recall	F-measure
nominal phrase	48	54	51
do not use till age of	93	98	96
do not use for older than	96	99	98

Table 3: Weighted results for both corpora. Results include all slots (143 in case of medical record corpus and 3 for drug leaflet corpus). Results are rounded up to integers.

Corpus	Precision	Recall	F-measure
Medical record corpus	99	64	76
Drug leaflet corpus	55	49	52

For the drug leaflet corpus the extraction of nominal phrases was not as successful as identifying data in the medical record corpus. It is because the extraction of phrases is complicated by the fact that both conjunctions and punctuation marks separate either two independent nominal phrases (as shown on Figure 2) or two variants of the same nominal phrase – in this case it is a kind of ellipsis (shown in Figure 3). The decision of which case it is can be based for example on the knowledge of the sentence structure. But at present, there is no NLP tool for Czech that is able to gain such information from the processed text.

Discussion

Generally, we have concluded that our approach to the problem is valid. Therefore the development of a system ready to be deployed within a health care provider has been initiated.

The results are promising as they show that it is possible to automate a great amount of manual work needed in a process of transforming free medical texts to a structured form.

In the case of medical record corpus the result are very promising and the described approach can be used in the practical applications. The data gained as a result of non-supervised transformation of medical records are ready to be used e.g. for purpose of statistics. However, extracted data must be verified by a specialist before they are entered into patient's health record.

We believe that the recall ratio for the medical record corpus can be further increased by improving the quality of the input texts. Medical records are full of typing errors and physicians` private abbreviations. Another issue is a quality of annotations. It is important to annotate all the texts consistently, but it is hard to achieve it as it is done by a number of annotators.

On the other hand the results of drug leaflet corpus processing are surprisingly low. It shows that such organised corpus does not provide enough information for creating good extraction rules. That is quite surprising, as we have believed that linguistic tags would provide enough information about the structure of the sentences and nominal phrases. But AMILCARE is not able to employ tags in such a way. This is obvious when looking at the set of extraction rules produced by the AMILCARE system in the stage of learning – linguistic tags are seldom used in the rules.

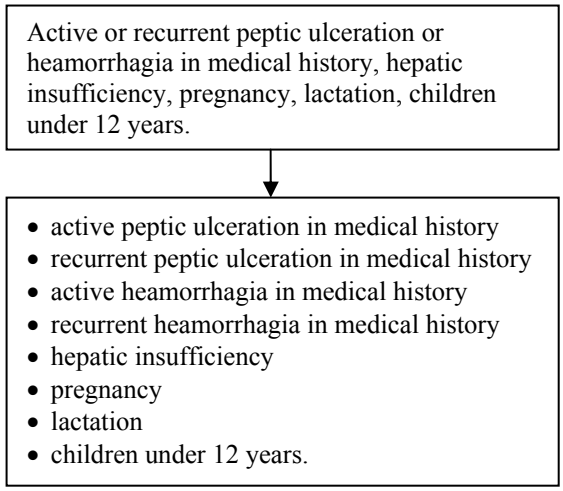

Figure 2: Example of compound nominal phrase. Comma separates independent nominal phrases. COXTRAL drug leaflet (Zentiva).

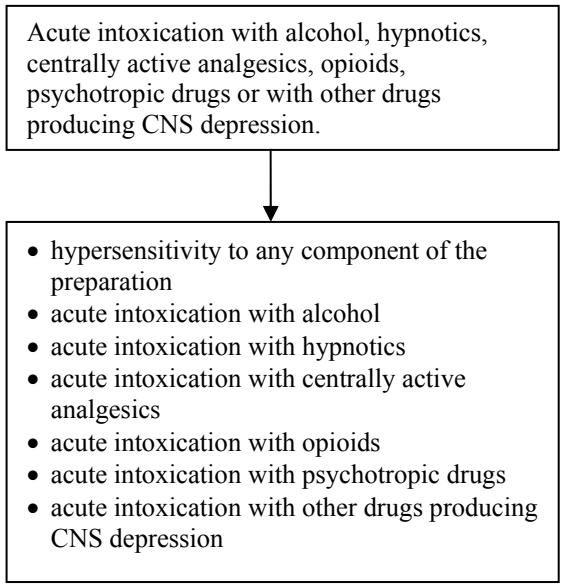

Figure 3: Comma separates variants of one nominal phrase. TRALGIT drug leaflet (Zentiva).

Another issue is that neither punctuation marks nor conjunctions do provide useful information, as they separate both independent nominal phrases and variations of a phrase. The results on a testing corpus show that many of the compound phrases are matched

ISSN: 1727-1983

partially (and thus decrease the recall rate). The low precision rate is caused by partial matches as well – the partial match is counted as wrong match.

We see three potential area of improvement: Extraction rule generation improvement – this is a potential problem as it implies substantial changes to AMILCARE system. The second area is to utilize further NLP such as sentence parser – this is problematic as well as it no off-shelf tool for Czech is available. The third possible improvement is to deliver better domain specific knowledge such as vocabularies and thesauri.

Conclusions

In this paper we have discussed the problem of transformation of medical text into structured form. We develop a system that should assist the experts in this activity by applying extraction rules to the free medical texts. As the means for building such system we use techniques and tools from the field of IE and NLP. We have also discussed several specific problems that arise when these systems are applied to the free text medical records and drug leaflets. One of the main aims of this project is to analyse and possibly overcome these obstacles and successfully apply the IE tools. Although we are currently only in the initial phase of the experiments we have reported the preliminary results with few basic insights. We believe that a deeper analysis of these results will allow us to modify the various tools and parameters involved in the training process such that the performance will rise. Apart from finishing the experiments and building the full system, in the future we would like to explore the possibility of enabling the training of the system also during its deployment, thus allowing it to adapt to a particular user.

The approach described in this paper can be used two fold. Firstly, it can be used for an automatic knowledge acquiring from patient records. This can be done without any human interaction. But the data gained does not meet the high requirements of precision necessary for human medicine. It can be used in area where the best-effort strategy is sufficient. The second use case is in the opposite edge: Its aim is not to automatically convert existing medical texts into computer-interpretable form, but to use NLP and IE technologies in order to enhance the productivity of already existing processes, such as manual transforming of medical knowledge into computer-interpretable form.

Acknowledgments

The work was supported by the grant number 1ET200300413 of the Academy of Sciences of the Czech Republic. The work was also partly supported by the Institutional Research Plan AVOZ10300504.

References

[1] FREITAG D. (1998): 'Multistrategy learning for information extraction', in: *Proceedings 15th International Conf. on Machine Learning*, pp. 161-169.

[2] CALIFF M., MOONEY. (1999): 'Relational learning of pattern match rules for information extraction', *Working Papers of the ACL-97 Workshop in Natural Language Learning*, pp. 9-15.

[3] SODERLAND S. (1999): 'Learning Information Extraction Rules for Semi-structured and Free Text', in Machine Learning.

[4] CIRAVEGNA F. (2001): '(LP)2, an Adaptive Algorithm for Information Extraction from Web-related Texts', in: *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, held in conjunction with *the 17th International Conference on Artificial Intelligence (IJCAI-01)*, Seattle.

[5] CIRAVEGNA F. (2001): 'Adaptive Information Extraction from Text by Rule Induction and Generalization', in: *Proceedings of 17th International Join Conference on Artificial Intelligence (IJCAI 2001)*, Seattle.

[6] AMILCARE Internet site address: http://nlp.shef.ac.uk/amilcare/lp2.html

[7] HAJIC J.: 'Disambiguation of Rich Inflection - Computational Morphology of Czech', Charles University Press - Karolinum, in press

[8] CAMPBELL DA., JOHNSON SB. (2001): 'Comparing syntactic complexity in medical and non-medical corpora', in: SUZANNE BAKKEN (Ed.), AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association, pp 90–94, Hanley & Belfus , Philadelphia.

ISSN: 1727-1983