FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Information Extraction of medical practice guidelines/protocols

State of the Art report for the course

188.948 Seminar for Medical Informatics

For the master studies

Medical Informatics (066 936)

Submitted by

Christian Hinterer
Matriculation-number 0927843

At the
Faculty of Informatics at the University of Technology Vienna

Supervision
Supervising tutor: Mag. Dr.rer.soc.oec. Katharina Kaiser

Vienna, November 29th, 2012

Table of Contents

Abbreviations

BPMN – Business Process Model and Notation

GATE – General Architecture for Text Engineering

GUI – Graphical User Interface

IDE – Integrated Development Environment

NLM – U.S. National Library of Medicine

UMLS – Unified Medical Language System

XML – Extensible Markup Language

IE – Information Extraction

MPG – Medical Practice Guidelines

CPG – Clinical Practice Guidelines

NLP – Natural Language Processing

SVM – Support Vector Machine

ICD – International Statistical Classification of Dieseases and Related Health Problems

CPT – Current Procedural Terminology

MeSH – Medical Subject Headings

SNOMED CT – Systematized Nomenclature of Medicine – Clinical Terms

LOINC – Logical Observation Identifiers, Names and Codes

GALEN – General Architecture for Languages, Encyclopedias and Nomenclatures in Medicine

EHR – Electronic Health Records

API – Application Programming Interface

WSD – Word Sense Disambiguation

KE – Knowledge Engineering

ML – Machine Learning

MHB – Many Headed Bridge

MUC – Message Understanding Conference

abstract

Thus to the complexity of medical therapies and their effective planning, medical practice guidelines (MPG) are even more important than in the past. The medical/clinical practice guidelines are in a free text format at the moment. This makes it impossible to process them in computer-based automated systems. Therefore some new languages were created that follow a formal and conceptual design and are structured in a pre-defined manor. Most of them use XML as their machine-understandable format. The key task is now to create concepts and applications that allow an easy modeling of CPG in these formats as well as translating the “old” free-text guidelines into the “new” designed formats.

To reach these goals of medical language processing, information retrieval and information extraction has to be used, as well as optimized and enhanced for the medical domain. To identify medical concepts in free text, algorithms have to make use of medical terminology systems. One famous terminology system is the UMLS, which contains over 100 known medical vocabularies/dictionaries. It combines a metathesaurus, a semantic network and a specialist lexicon.

Aim for the future is the creation and/or the improvement of existing medical language processing systems to create comfortable, user-friendly and seamless applications that offer functionalities for modeling medical knowledge, make it useable for other systems, make it 100% understandable for machines, offer import of existing unstructured medical documents as well as easy expandability.

Another aim would be that these applications help to close the gap between the medical domain and the formal domain of informatics.

1. Introduction

Medical therapy planning is growing to be more complex than ever before. The load of knowledge about diseases, affected patients, different intentions and the different conditions of each case make it even more complex. Therefore medical practice guidelines (MPG) were created to support this process of therapy planning. At the current state these guidelines are not available for automated computer-based processing. Therefore the Asgaard/Adbru[1] project was started. One Aim of the Asgaard/Asbru project is the development and the offering of tools that simplify the construction process of these complex therapy plans. [1, Miksch, 1997]

MPG contain information about medical processes, medical therapies and treatments, including lots of detailed information belonging to the diseases, conditions, the patients, the different branches of a therapy in dependency of different conditions and many more information belonging to a therapy process. This report contains state-of-the-art knowledge about how information could possibly be extracted of existing MPG and be imported into the Asgaard project, respectively translated into the Asbru language. To be more specific the report deals with the process of finding specific information, such as condition expressions, in the MPG with methods that belong to the process of information extraction.

The main challenge of this task is to find and even more to interpret the demanded information of the condition expressions that are contained in the MPG. The key to the solution is the correct interpretation of the found conditions, so that they can be translated into the correct, corresponding Asbru conditions.
Therefore metathesauri are used that contain the medical concepts, this makes the finding and identification of the information possible in the first place. The metathesauri deliver the information that is used as the domain of the information extraction process.

[1] Asgaard/Asbru is a project of the Vienna University of Technology and the Stanford medical informatics: http://www.asgaard.tuwien.ac.at/about/project.html [November, 2012] [11]

2. Related Work

This chapter contains an overview of the existing projects and their executing communities as well as existing literature that deals with the topics of Information Extraction (IE) in general, IE out of MPG, interpretation of information contained in condition expressions, interpretation of medical texts and statements and translation of information into the Asbru language and metathesauri for medical concepts and their sense and usage.

- Clinical practice guidelines [Fletcher, 2007]
  Description of Clinical practice guidelines and their determined usage.
- Learning Information Extraction Rules for Semi-Structured and Free Text [Soderland, 1999]
  Stephen Soderland, of the University of Washington, Seattle, describes how important IE become due to the mass of available on-line text information and the need to automated process this information. He says that each IE application needs a separate set of rules tuned to the information domain to function properly. He also describes the use of whisk, a general rule extraction system. He also describes how free text can be searched for demanded information by using the methods of syntactic analysis and semantic tagging along with the whisk rule extraction system.
- Ontologies and Information Extraction: A Necessary Symbiosis [Nédellec and Nazarenko, 2003]
  The authors describe IE including the process of it, ontologies, how ontologies are used for IE by identifying the relevant information and how information extraction is used to design and/or enrich ontologies.
- Information Extraction from World Wide Web [Eikvil, 1999]
  A Description of IE for the world wide web, which means IE of structured and semi-structured documents, the use of wrappers to achieve IE, and wrapper tools and how they work.
- Extraction Patterns for Information Extraction: A Survey [Muslea, 1999]
  A Survey of existing IE patterns to compare their powerfulness.
- A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text [Chieu and Ng, 2002]
  Description of a classification-based approach for IE using an algorithm to determine the maximum entropy of information.
- SVM Based Learning System For Information Extraction [Li, Bontcheva and Cunningham, 2004] (community of the GATE)
  Guidance of developing a “support vector machine” for IE and a comparison to other approaches/methods.

- Bio-medical entity extraction using support vector machines
  [Takeuchi and Collier, 2005]
- Gaining Process Information from Clinical Practice Guidelines Using Information Extraction [Kaiser, Akkaya and Miksch, 2005]
  A heuristic approach for Information Extraction of Medical practice guidelines.
- Supporting the Abstraction of Clinical Practice Guidelines Using Information Extraction [Kaiser and Miksch, XXXX]
  Using NLP methods to provide helpful auto-generated information. Using a rule based system on syntactic and semantic information and semantic relationships from the UMLS Semantic Network. Analyzing the MPGs mainly for Conditions and Actions.
- Information Extraction Approach for Clinical Practice Guidelines Representation in a Medical Decision Support System [Pech-May, Arecalo and Sosa-Sosa, 2011]
  Importance of MPGs and even bigger importance of analyzing them to translate all the contained information into a format that can be interpreted and processed by computer systems. This paper deals with a possible approach to get the most effort of information out of the guidelines while translating them into a structured format. Description of a 3-level approach to structure MPGs, including condition expressions. This paper also includes a possible translation of information into the Asbru language.
- Sentiment Analysis of Conditional Scentences [Narayanan, Liu and Choudhary, 2009]
  Making use of Sentiment Analysis to split conditions into three main groups, namely positive, negative and neutral ones. This makes it possible to classify the conditions in an even more precise way. The degree of fineness of the realized classification determines the complexity of the process that shall identify the semantics, respectively the contained information of the conditions.
- Using the Unified Medical Language System (UMLS) Terminology Services to abstract medical information out of unstructured texts.
  [https://uts.nlm.nih.gov/home.html, December, 2012]
- A study of PROforma, a development methodology for clinical procedures [Vollegregt, Teije and Harmelan, 1999]
  Designing medical processes with the PROforma methodology. A graphical process plan including, sequences, decisions, actions and enquiries
- Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program [Aronson, 2001]
  A basic illustration of the MetaMap implementation concept, an application to find biomedical concepts in texts with the usage of the metathesaurus of the UMLS.

- The Unified Medical Language System (UMLS): integrating biomedical terminology [Bodenreider, 2004]
  An introduction to the UMLS and an explanation of the idea of the UMLS as well as an example for the usage and an introduction to the related tools.
- Medical practice guidelines and protocols: the Asgaard/Asbru project [Miksch, Sahar and Johnson, 1997]
  Introduction to the Asgaard/Asbru project and explanation for the need of it.
- Identifying Actions Describes in Clinical Practice Guidelines Using Semantic Relations [Kaiser, Seyfang and Miksch, XXXX]
  Description of the methodology used to identify “actions” as a medical concept in CPG using the UMLS metathesaurus within the application MetaMap to identify semantic relations.
- Medical Terminology Systems [Kaiser, 2007]
  An Overview of medical terminology systems its purposes and existing coding systems for the abstraction of medical records aiming the goals of completeness, semantics, structure and system integration.

3. Method

All of the described results in this report are originate from scientific papers, articles of academic journals and other subject-specific expert literature.
We were searching the web directly, using “Google” and “Google Scholar”. We were searching the publisher “Springer” for information, using the “Springer Link”.
We were also was searching the “ACM Digital Library” and the “IEEExplore Digital Library” for Information, but due to access control, we could not get the full information range we were looking for. A very useful search engine for my researches was the “PubMed” of the US National Library of Medicine (NLM).

The read papers and articles are listed in chapter 2 (Related Work). Note that not all of them are cited in this report.

In the process of relevant information finding, we were using the following search terms:

- information extraction
- medical practice guidelines
- information extraction of medical practice guidelines
- [information extraction] combined with [unstructured, semi-structured, structured] [sources, documents, information]
- information extraction tools
- formalization of medical knowledge
- Asgaard
- Asbru [language]
- sentiment analysis
- information extraction condition finding
- information extraction state of the art
- medical information extraction state of the art
- unified medical language system
- medical thesaurus
- medical language processing

While the number of overall results was innumerable, the results regarding the recent past were very rare.

To filter the huge amount of found papers, reports, articles and other literature for useful information was a very time expensive and partly iterative process.

The selection criteria for the literature used in this report was to find information, respectively knowledge that deals mainly with the report topic, namely the information extraction and exploitation of medical knowledge in documents that are unstructured, semi-structured as well as structured and their further processing.

The selection of the relevant results was done with the following (iterative) process:

- searching for literature using the search engines and keywords described earlier in this chapter
- reading the abstract and the table of contents
- using the document search function with the keywords
- browse the relevant chapters
- compare the literature with already found and evaluated literature
- trying to assess the quality of the literature
- assess the grade of relevance -> decide for the usage: if the literature was important for this report, we downloaded it, categorized it, assigned it to one of the groups: information extraction in general, information extraction of medical information, automated technologies for information extraction, tools for information extraction, medical language systems/thesauri, medical language processing; (most of the found literature fits at least two groups) and at end we appended title, authors, publication year and rough content to my list.

Literature that was excluded was either already out of date, which means the publication date was much older than 10 years, or contained mainly the same content that was already included in other, already used, literature.

4. Results

4.1 Information Extraction

Information Extraction is the process of parsing free-text, semi-structured text or structured text for information of interest and extracting the information, mostly with the intention to further process the found information.
Since this is a very complex process and therefore very difficult to implement in an automated system, some different approaches where developed in the past. The next chapters will illustrate some approaches with different methods and partly different sources, as well as aims. The key task of information extraction is the identification of the relevant information. Therefore a domain is needed, that defines the concept that shall be identified. This domain could possibly be a dictionary, an ontology, a thesaurus or even better a metathesaurus. In (pre)semi-/structured texts, the relevant information could possibly be already tagged on its creation, which makes the process of information extraction much easier. For example if the relevant information is already tagged with an own xml-tag and the consumer of this xml-document knows the schema, it is only necessary to pick the relevant xml-tags out of the document. The text within the xml-tag still has to be analyzed, but at least it can be found very quickly, without the need of complex and high-cost search algorithms. [3] [4]

4.1.1 Information Extraction and Ontologies

A possible way to provide the domain for information extraction is to use an ontology that represents the domain. An ontology is a description of conceptual knowledge organized in a computer-based system. Using ontologies for IE is a cyclic process. On one hand the ontology is used as domain for IE to identify the relevant information and on the other hand the extracted information is used to enhance the ontology. The ontology of course can only deliver the semantic knowledge for the IE process, lexical knowledge and grammar that defines the syntax is needed additional.

There are 4 levels of ontological knowledge to distinguish regarding IE:

- Referential domain entities and their variations are listed in “flat” ontologies. That means ontologies with flat structure. The main usage is identification and semantic tagging of string (specific words of the domain) in documents. The following three objectives can be distinguished, although these operations are usually processed together.

- Semantic tagging: Lists of entities are used to tag the text entities with the relevant semantic information. In ontologies, entities are described by their type. Exact character strings are not the best way of entity identification and semantic tagging, because they are not precise enough. In many languages one word has different meanings, and the actual meaning results of the context of the sentence. Therefore it is necessary to define contextual IE rules.
- Naming normalization: The resources of semantic tagging are used for normalization. This means that various forms of the same semantic entity are tagged as the same entity and associated to the entity type. Especially genomic systems are concerned with the variation problem, because the nomenclatures are not consistent in literature.
- Linguistic normalization: This solves some syntactic ambiguities. For example if “cotA” is tagged as a gene, in the sentence “the stimulation of the expression of cotA”, knowing that a gene can be “expressed” helps to understand that “cotA” is the patient of the expression rather than its agent or the agent of the stimulation action.

- The conceptual hierarchy improves normalization by enabling general levels of representation. Besides lists of entities, ontologies often described as hierarchies of semantic or word classes. IE makes use of word classes rather than hierarchical organization for semantic tagging. Medical language systems learn extraction rules by generalizing from annotated training examples.
- Some IE systems only use chunks of their domain model. These chunks contain descriptions for properties and interrelations of entities. The projection of these relations improves the IE process. In this case expression rules ensure the mapping between a conceptual node of the domain and the potential expression that contains the relevant information.
- The domain model concept itself is used for inference reasoning between syntactic level and event description.

Information extraction using ontologies as domain manly happens in two steps. First step is to identify the phrases or words in the text that comply the ontology’s entities. In the second step the relations between the entities are identified and extracted. [3]

4.1.2 Information Extraction from the World Wide Web

The www consist of many online documents. The facts that online documents have very large volumes, that their content changes often, that often new documents appear, that they often include much semi-structured text and that they often include hyperlinks make it even harder to extract information out of them.
Therefore IE from the web is often performed using wrappers. In this context a wrapper can be seen as a procedure that extracts content of a given document and

delivers the relevant information in a self-describing representation. The wrapper consists of a set of extraction rules and the logic that applies these rules to a specific information source, often called the corpus. The wrapper generation deploys techniques that are less dependent on full grammatical sentences compared to usual NLP techniques. Thus to the often changing content of a web document research has led to automatic creation of wrappers.

The different wrapper systems are listed here just for completeness and can be researched in other literature on further interest: ShoptBot, WIEN, SoftMealy, STALKER, RAPIER, SRV, WHISK; [4]

4.1.3 Information Extraction using SVM based learning systems

Due to the problem that found entities in the text are often consisting of more than one word, a new system to deal with this problem had to be designed. The use of classifier-based IE systems solves this problem. For example a two svm classifier system can be used. One classifier is used to recognize the beginning on an entity and another one for the end. Another two classifiers could be used for the middle words and for single-word-entities. Or another classifier to find words that does not belong to a defined entity. Due to the use of more than one classifier to tag a word or a phrase, some post-work is needed. After the tagging, each multi-tagged word or phrase has to be assigned to one single tag. Therefore first of all, all start- and end-tags has to be removed that does not have a corresponding tag. Then all possible tags for one word hast to be compared to choose the best one out of it. [5]

4.2 Medical Language Processing

Clinical practice guidelines are documents in narrative form, therefore they are often ambiguous and lack of a defined structure and internal consistency. This makes it impossible to process them directly by a computer without pre-processing. To make them understandable for a computer, the usage of a formal representation language is necessary. There are many existing different representation languages, frameworks, approaches and also tools to model CPG. Translating CPG, into these formal languages that can be interpreted by computer-systems, manually, as well as designing new CPG is a complex and time-consuming task that requires both, knowledge about formal methods and about the medical domain. [8] [9]

4.2.1 Clinical/Medical Practice Guidelines

CPG are documents that contain recommendations for health care personal about care and treatment of patients. The more the CPG is based on research evidence the higher the quality of the guideline. The two parts of a guideline are the clinical question with the strength of the evidence that affects the clinical condition-making for that condition and the set of recommendations for the management of each patient under different circumstances, respectively conditions.
CPG are mainly for the usage by clinical personal, but are also used by insurers and administrators of hospitals. Its main advantages shall be quality improvement of therapy and therapy planning and the reduction of costs.

The base of each guideline should be an evidence research, though expert opinions can also be involved in guidelines, but should be marked as such. Due to the fact that the guidelines are results of researches and collected knowledge, the creation and improvement of them is an iterative process.

Guidelines should consider the magnitude of effect, harms from the intervention, convenience and side effects, clinical skills that are needed to execute the tasks, contained in the guideline, properly and patient records, but also costs and cost-effectiveness.

A quality check of guidelines is done by an expert group, charging on 18 different conditions.

Although studies proved the positive effects of guidelines in practice, they also showed up that there are just small positive effects. [2]

4.2.2 Formal representation languages for CPG

As already mentioned, to make the CPG machine-readable they have to be translated into a formal language. Most of the existing formal languages use XML as their format. The following list contains some existing formal languages:

- Asbru: A task-specific and intention-based plan representation language, designed for management-task plans.
- GLIF (Guidline Interchange Format): The GLIF defines an ontology for the representation of the CPG, as well as it includes a formal expression language for specification of decision criteria and patient state.
- GEM (Guideline Elements Model): This is also an XML-based guideline document model that contains the CPG information in an organized way.
- EON: EON is a modeling as well as execution system for CPG and part of the EON architecture, a component-based suite of models and software components for the implementation of CPG-based applications.

- PROforma: In PROforma the CPG can be modeled as a set of tasks and data items. It is designed to support the medical procedures and the decision making at the point of care. It splits all tasks into four groups: plans, decisions, actions, inquiries.

Following concepts are included in these languages:

- Plan organization: Abru and PROforma use an isolate gnereic object class for modeling plans; GLIF uses two types of plans, namely guides and macros.
- Specification of goals/intentions: Asbru represents the intentions of plans as temporal patterns depending on the context.
- Action model: Actions in the formal languages represent the tasks contained in the CPG. In Asbru the effects of actions can be used to choose between different alternative plans and express causal relations.

There are several existing tools for modeling CPG in these formal languages. They are just listed for completeness, because this report does not deal with the creation of new CPG, but with the translation of existing ones into a formal language: [8]

- Stepper
- GEM Cutter
- DELT/A
- Uruz
- Protégé
- AsbruView

4.2.3 Medical Terminology Systems

Medical Terminology Systems are used as domain in the process of information extraction of medical documents such as MPG.

Due to the complexity and comprehensiveness of medical knowledge, it is necessary to organize it in medical terminology systems. Possible concepts are a nomenclature, which is a simple system of names. A vocabulary is a system of names including explanations and meanings. A classification systematically organizes the knowledge into classes and a thesaurus is an index of medical literature and offers support for searching over bibliographic databases. The importance is not only the knowledge itself, but also how the knowledge is stored and handled.

Coding Systems for medical records have been developed to optimize the handling of medical data records. One of the first collections of medical data was the ICD. The first ICD revision consisted diagnosis organized with codes. Later revisions also included other medical knowledge such as procedures. Revision number ten categorized the existing codes.

The CPT is a nomenclature used to report medical procedures and services. It is still in use, for example in the united states for billing.

MeSH is a thesaurus containing medical vocabulary for the use of indexing, cataloging and searching of health-related information.

Since the use of EHR (electronic health records) there was a need for other coding systems, because the already mentioned ones could not fulfill the requirements for being implemented within EMR systems. The main reason is that these coding systems are not detailed and precise enough. Thus, the first EMR systems developed and used their own coding system.

As a result of the importance of terminologies, the demand for reusable and shareable multipurpose systems was big. Theses “new” terminology systems should support “classification and coding systems”, EMR, “decision support systems”, “knowledge management systems” as well as “natural language processing”.

SNOMED for example combines a coding system, a vocabulary, a classification system and a thesaurus. It is designed to handle patient’s history, diseases, treatments and outcomes.

LOINC is a system developed to manage codes for laboratory test results and other clinical observations. LOINC is compatible with HL7, a group of international standards for the communication of health care systems. [6]

The UMLS is a collection of many international vocabularies and classifications and provides a mapping structure between them. It is build up of three knowledge components. The first one is the metathesaurus, which contains more than 100 vocabularies including all of the so far mentioned in this report. This large number of information sources in different formats, following different standards, made it necessary to map between them, to keep the information useful and in the right context. The second component is the semantic network, which is an ontology providing an overarching framework for all other UMLS concepts. It defines categories in which the concepts of the metathesaurus are categorized and semantic relationships that are assigned between these concepts. The third component is the specialist lexicon, which contains syntactic, morphological and orthographic information for biomedical and other words in the English language.

Thus to the enormous number of different information sources combined in the UMLS, the processing time of queries would be too long and the number of results would be way to enormous and also including likewise very similar results due to similar information in the sources. Therefore the MetamorphoSYS can be used to configure UMLS metathesaurus regarding language restrictions and many other options.

Due to the enormous trunk of information provided by the UMLS by combining many prominent medical information sources, it is recommended to use the UMLS when implementing information systems, such as information extraction applications, to identify all the medical concepts and to deal with the input texts from many different angles. [7]

4.2.4 Information Extraction of Clinical Practice Guidelines

In the process of IE of CPG the CPG are the corpus, respectively the information source. The domain knowledge, respectively the medical concepts has to be given by an information source for medical concepts and knowledge. The recommendation for the domain is the UMLS (chapter 4.2.3) because it provides an enormous amount of medical concepts and information by combining over 100 known different medical vocabularies and it also provides API and plug-ins that make it available when building your own automated computer application for processing CPG.

The challenge of IE of CPG is to identify and link together all the relevant medical concepts, while ignoring the irrelevant information contained in the source document. The two different approaches IE can be classified are Knowledge Engineering (KE) and Machine Learning (ML). While KE is based on a domain corpus and focuses on an empiric method to develop NLP systems, ML has a well-known set of documents and outputs and uses patterns to extract knowledge. The ML approach can support the IE process to extract information from CPG for enhancing its formalization.

One possible approach is to create templates that are created by a specific xml schema. Next the CPG are parsed for relevant medical concepts such as actions, processes, sequences, conditions etc. and the information is written into the templates.

The key task is to identify the relevant medical concepts and their semantics. Once it is known which information could be gained from the CPG, it is relatively simple to translate it into a structured format like XML.

One fact that makes it hard to implement computer-based modeling systems for CPG is that the developer needs both, skills with modeling methods and techniques, as well as detailed domain model knowledge, which is often very complex and complicated regarding to an extraordinary domain like the medical one.

An existing format is the many-headed-bridge (MHB). This format is a XML language for abstracting a CPG while translating it into another language. MBH provides chunks of information which correspond to chunks of the natural language text of the CPG. These chunks are split into 8 different groups that cover the medical concepts. The 8 groups are:

- control flow: execution order of tasks in the guideline, their decomposition and the gathering of information; it can be used to specify decisions, ordering, decomposition and synchronization of tasks and actions
- data flow: data processing included in treatment
- temporal aspects: qualitative as well as quantitative
- background information
- evidence
- resources

- patient related aspects
- document structure

The process of identifying information and medical concepts starts with annotating nouns and phrases that include these nouns. Then the annotations have to be mapped to the corresponding medical concepts. In most cases annotations fit to more than one medical concept. Depending on the syntactic terms and the semantic relations, the relevant medical concept that fits the annotation best is chosen. Then the chosen medical concept is linked to related concepts to recreate logical relations that were represented in the CPG. Beginning at the most inner concepts the whole structure of the semantics and the relations of the concepts is recreated in a structured form. The schema of the structured document is determined by the target system that further processes the information.

The NLM already provides some applications that at least identify the medical concepts in free texts in some different ways. This provides a profound basis to start with processes and/or application building to further analyze and process the information in question. [8] [9]

One of these applications is MetaMap, described in the next chapter.

4.2.5 MetaMap to identify medical concepts in CPG

MetaMap is a program developed and provided by the National Library of Medicine to find medical concepts in free text by using the UMLS metathesaurus, respectively mapping free text phrases to medical concepts of the metathesaurus. MetaMap is high configurable and free available for everyone who registers at the NLM (http://nls9.nlm.nih.gov).
MetaMap explores input text for medical concepts given by the metathesaurus. It parses the text and creates some variants of the found medical concepts and their relations. It creates candidates out of the variants and evaluates them by first computing a mapping from the phrase words to the candidate's words and then calculating the strength of the mapping using a linguistically principled evaluation function consisting of a weighted average of four metrics: centrality (involvement of the head), *variation* (an average of inverse distance scores), *coverage* and *cohesiveness*. The latter two components measure how much of a candidate matches the text and in how many pieces. The candidates are then ordered according to mapping strength. [Aronson, 2001]
Then complete mappings are constructed by combining the candidates involved in one phrase and then the strength is computed for the phrase just like for candidate mapping. The highest score represents MetaMap's best interpretation for the original phrase.

MetaMap can not only be used online for direct processing, it can also be downloaded and run local on your computer.
It also offers an API for integrating it in your own JAVA implementations.
The most useful part for application programmers is the plug-in for GATE. This allows the developer to use the MetaMap to find medical concepts in free text and at the same time using the results of the MetaMap analysis for further processing using the GATE functionality. [10]

5. Discussion

In this section we will give an overview of the advantages and disadvantages of the different approaches and results contained in chapter 4 (Results).

First thing to explain is the importance of machine learning in acquisition of text extraction. Although no automated information extraction system reaches the same quality of human competence, it is very important to implement automated machine learning text extraction systems. The reason is that the effort and expertise required for manually constructing extraction rules is pretty high and this makes it very hard for people who lack one of both, either knowledge of the domain or knowledge of extraction systems.

Information extraction systems in general are very knowledge-intensive and require extraction rules for the domain of which the information is to be extracted. Among others, this is one reason that makes machine learning very attractive for this process, instead of creating these rules all by hand. Without rules based on the domain, it would be impossible do identify the searched concepts in an input text.

The simplest text to extract information from is structured text. This has the advantage that the information is already tagged and therefore it can be found rather easy, compared with unstructured text. Semi-structured text contains the relevant information in a fairly small number of stereotyped contexts. Free text is the most challenging sort of text to extract information from. When extracting from free text, it is necessary to execute some syntactic analysis and semantic tagging before using extraction rules. [12]

Information extraction using ontologies to provide domain knowledge is abundant and multiple. There are many existing different approaches, algorithms and systems. Since there is no general approach that gives the best precision and recall it is difficult to evaluate which of the different approaches rules the others. The main influences on the evaluation are statistics of natural language processing, the influence of the message understanding conference on IE and the cost of ontological processing. But the growing needs for better and better automated information extraction systems will result in a stronger involvement of ontological knowledge. The main problem is the discrepancy between what the relevant text is about, the exogenous lexicon and the ontology. To provide better ontology-based extraction systems this gap has to be closed. [3]

When using information extraction on web pages the usage of wrappers is essential. The construction of these wrappers is often tedious and requires expert knowledge. Since there are lots of dynamic web pages the costs of wrapper maintenance are very high. As a result automatic wrapper construction has become a focused problem comparable to the automatic rule creation on rule based machine learning IE systems. To solve or at least to improve this problem there are already several

(academic) implemented systems like ShopBot, WIEN and STALKER for structured sites and RAPIER, WHISK and SRV for less structured sites.

Nowadays search engines are able to return lists of found documents, but they are not powerful enough to already extract the relevant information from these documents. Hence XML defines the content rather than the presentation, it makes wrapper induction simpler, but it will not eliminate the need for wrappers. [4]

Using SVM-based IE systems shows that they are comparable to other state of the art IE systems, but tend to be more complex, especially for big information input. When using SVM-based IE systems it showed that treating the negative and the positive margins not equally, the results are much more efficient compared to even margins. The main disadvantage of using SVM-based algorithms is that every entity type needs several SVM classifiers which make the approach even with only a few entity types quite fast quite complex. [5]

Since clinical practice guidelines are documents in narrative form, they are lacking defined structure and internal consistency. This makes it impossible to process them directly by a computer without pre-processing. That is done by putting them into a more formal representation. One aim of the future is to flip the situation, namely to store and process CPG in a computer-processable format and to prepare them for presentation and not to create them as plain text and as a result deal with the complex task of getting the contained information into a computer system. [2] [8]

To process these formal documents, systems were developed that understand the content of these documents, so called medical terminology systems. The first systems that were designed were mainly to support administration of medical facilities such as hospitals. These systems have the big disadvantage that they do not cover all diagnosis codes. Other systems only cover a small part of the processes in medical facilities. Such systems are often proprietary, limited, custom-built and difficult to use due to low usability. As a result some of them have quite low user acceptance. Although many of the existing medical vocabularies overlap, they cover different parts of the medical language which makes the corresponding coding systems difficult to compare. This also makes interoperability to a significant problem. To integrate patient data with health information technologies comprehensive clinical terminology systems are needed. [6]

A medical terminology system that covers many different international vocabularies and classifications is the UMLS. It consists of three different knowledge components and contains over 100 different vocabularies. With the use of the MetamorphoSYS it is highly configurable, to prevent getting to much (duplicate) results when querying it. Since the UMLS is free, high configurable, contains lots of different vocabularies and also includes different tools to use and to query it, it is recommended as underlying terminology system, when implementing information systems for the medical sector. [7]

One of the included tools is the MetaMap, which is an effective tool for identifying medical concepts in documents with the help of the UMLS metathesaurus. The

MetaMap tool has two disadvantages. Firstly it has problems with detection of idiosyncratic text, acronyms and abbreviations and numeric quantities. Secondly it has problems with the resolution of ambiguity. The first problem is solved by using an extensible, hierarchical tokenization regime. The problem of ambiguity is solved by using WSD methods. [6]

When extracting information out of CPG the main challenge is to have skills and knowledge of the medical terminology used in the CPG and of information extraction itself. When creating computer-based solutions the biggest problem is the complexity of the knowledge and information contained in the extraordinary medical domain. Many developers have knowledge of modeling techniques, but no medical domain knowledge. The main intention when extracting information of CPG is to transform the relevant information into a format that can be processed by computer systems. Although there are already a few different approaches to extract information or at least some special types of information out of CPG, there is no existing solution that can identify and extract all the relevant medical concepts and its relations. One promising approach is developed at the University of Technology in Vienna, where different methods for modeling and extraction are used and the UMLS as domain model for the medical knowledge. [8] [9]

6. Conclusion

Our research came up with the result that there are already many different approaches, methods and tools for computer-based automated information extraction. All of the existing approaches have one thing in common: They need some kind of information source that provides them the concepts and knowledge of the domain model and some rules how the relevant information should be extracted. The domain model knowledge could be provided as some kind of dictionary, ontology or even matathesaurus. Although there are existing machine learning rule based systems, they do not reach the efficiency of correctness like done by humans manually.
The compromise is to use automated information extraction with manual rule definition. This again is a very complex task itself, because the operator of this task needs both, expertise knowledge of modeling techniques and expertise knowledge of the language domain, which is a very complex one talking of the medical knowledge domain.

Although there are existing approaches for automated medical language processing of medical documents, there is no seamless implementation that is able to process all medical concepts, like for example provided by the UMLS, in all kinds of medical documents, like for example CPG.

Aim for the future is to improve existing and to implement new solutions that accept configurations for the used domain model and for the input documents and provide structured and standardized results that contain all relevant information, knowledge and its relations. The exact aim would be to define extraction rules dignified enough to extract only exactly that information that it needed and this in every different use case. But since there are still so many different formats, standards and systems, this is – at the current state of the art – impossible.

Bibliography

[1] Medical practice guidelines and protocols: the Asgaard/Asbru project [Miksch, Sahar and Johnson, 1997]

[2] Clinical practice guidelines [Robert H Fletcher, MD, MSc., 2007, version 16]

[3] Ontologies and Information Extraction [Nédellec and Nazarenko, 2003]

[4] Information Extraction from the World Wide Web [Eikvil, 1999]

[5] SVM Based Learning System For Information Extraction [Li, Bontcheva and Cunningham, 2004]

[6] Medical Terminology Systems [Kaiser, 2007]

[7] The Unified Medical Language System (UMLS): integrating biomedical terminology [Bodenreider, 2004]

[8] Information Extraction Approach for Clinical Practice Guidelines Representation in a Medical Decision Support System [Pech-May, Arecalo and Sosa-Sosa, 2011]

[9] Supporting the Abstraction of Clinical Practice Guidelines Using Information Extraction [Kaiser and Miksch, XXXX]

[10] Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program [Aronson, 2001]

[11] The Asgaard/Asbru Project. Retrieved At: October 16, 2012. http://www.asgaard.tuwien.ac.at/about/project.html

[12] Learning Information Extraction Rules for Semi-Structured and Free Text [Soderland, 1999]